Universidade de Brasília - UnB

Engenharia Elétrica

Utilização de Aprendizado por Reforço para Operações em Bolsa de Valores

Autor: Stefano Giacomazzi Dantas

Orientador: Prof. Dr. Daniel Guerreiro e Silva

Brasília, DF

2017

Stefano Giacomazzi Dantas

Utilização de Aprendizado por Reforço para Operações em Bolsa de Valores

Monografia submetida ao curso de graduação em Engenharia Elétricada Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia Elétrica.

Universidade de Brasília - UnB

Orientador: Prof. Dr. Daniel Guerreiro e Silva

Brasília, DF

2017

Stefano Giacomazzi Dantas
Utilização de Aprendizado por Reforço para Operações em Bolsa de Valores/ Stefano Giacomazzi Dantas . – Brasília, DF, 2017-
73 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Daniel Guerreiro e Silva

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
, 2017.

1. Aprendizado por Reforço. 2. Bolsa de Valores. I. Prof. Dr. Daniel Guerreiro e Silva. II. Universidade de Brasília. III. IV. Utilização de Aprendizado por Reforço para Operações em Bolsa de Valores

CDU 02:141:005.6

Stefano Giacomazzi Dantas

Utilização de Aprendizado por Reforço para Operações em Bolsa de Valores

Monografia submetida ao curso de graduação em Engenharia Elétricada Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia Elétrica.

Trabalho aprovado. Brasília, DF, 11 de Dezembro de 2017:

Prof. Dr. Daniel Guerreiro e Silva
Orientador

Prof. Dr. Alexandre Romariz
Convidado 1

Prof. Dr. João Paulo Leite
Convidado 2

Brasília, DF
2017

Aos meus pais e irmã, que nunca deixaram faltar
apoio e carinho na minha vida.

Agradecimentos

Gostaria primeiramente de agradecer ao meu orientador, professor Daniel Guerreiro, que aceitou embarcar nessa "aventura"que é trabalhar com mercado de ações, suas orientações e conhecimento foram essenciais para completar esse trabalho. Gostaria também de agradecer ao professor João Paulo Leite, por me convencer a seguir com esse tema quando eu pensei em desistir, além de ser sempre solícito quando eu precisava de orientações sobre Telecomunicações. Ainda, gostaria de agradecer ao professor André Noll, que apesar de não ter participado desse trabalho, foi fundamental na minha caminhada até aqui. Não poderia deixar de agradecer aos meus amigos, em especial: André Costa, Arthur Carvalho, Gabriel Bayomi, Gabriel Castellano, Henrique Orefice, João Antônio Rondina, Leonardo Alburquerque, Letícia Brito, Luiz Felipe Campos, Pedro Campos, Renata Lopes e Gustavo Cid. Todas as conversas, os incentivos ("relaxa, vai dar certo") e as brincadeiras foram fundamentais em todos esses anos de Engenharia Elétrica, um curso que pode ser extremamente desgastante. Considerando tudo que passamos, a frase *"If you want to go fast, go alone. If you want to go far, go together"* nunca fez tanto sentido. Gostaria de agradecer às páginas CBM e SAM, que me fizeram rir em momentos onde o estresse reinava. Por último, mas não menos importante, gostaria de agradecer aos meus pais e irmã, vocês são a fundação onde minha vida está sendo construída.

“If it doesn’t challenge you,
it doesn’t change you.“
(Fred DeVito)

Resumo

Este trabalho tem como objetivo mostrar como um agente inteligente, treinado por meio do algoritmo *Q-Learning*, pode obter resultados consideráveis ao operar na bolsa de valores, usando dados financeiros ruidosos, não-lineares e não-estacionários. Para avaliar o desempenho do agente, construiu-se um simulador da bolsa de valores. Utilizando quatro indicadores técnicos como parâmetros de entrada e uma adaptação do retorno diário como função de recompensa, o agente obteve desempenho superior ao *Buy & Hold* em metade dos casos. Além disso, abre-se espaço para a discussão acerca do comportamento de ações de diferentes setores da economia norte-americana e das possíveis limitações da análise técnica, de acordo com os resultados obtidos.

Palavras-chaves: Aprendizado por Reforço, Bolsa de Valores, *Q-Learning*, Aprendizagem de Máquina

Abstract

This work aims to show how an intelligent agent trained by the Q-Learning algorithm can achieve intesting results, using non-linear, non-stationary and noisy financial data. In order to evaluate the system perfomance, a stock market simulator was developed. The agent outperformed the Buy & Hold strategy in half of cases, using technical indicators as input data and a modified version of daily return as the reward function. Furthermore, the stock market behavior of different economy sectors is discussed along with possible limitations of technical analysis.

Key-words: Reinforcement Learning, Stock Market, Q-Learning, Machine Learning

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Contextualização

Historicamente, o dinheiro tem extrema importância nas nossas escolhas, tanto profissionais como pessoais. Por esse motivo, a relação dos seres humanos com o dinheiro é estudada constantemente por áreas da psicologia. Nos estudos apresentados em (VOHS; MEAD; GOODE, 2008), mostra-se como o conceito de dinheiro provoca uma mudança nas nossas relações interpessoais. Já (VOHS; MEAD; GOODE, 2006) mostra o efeito que o dinheiro traz na motivação das pessoas.

Um dos lugares mais procurados para o enriquecimento é o mercado de ações. Esse ambiente é retratado com frequência na mídia, séries de TV e filmes como um cenário caótico e estressante. William Feather[1] disse "o engraçado do mercado de ações é que toda vez que alguém compra, outra pessoa vende, e os dois acham que fizeram um bom negócio". Não por acaso, sempre há risco de perder dinheiro, ainda mais se nenhuma estratégia for adotada pelo investidor. Com isso em mente, diversas técnicas foram desenvolvidas ao longo dos anos para obtenção de sucesso no mercado financeiro.

Segundo (MOODY; SAFFELL, 1999), o objetivo final de um investidor é otimizar alguma métrica, seja o lucro, o retorno ajustado pelo risco ou a utilidade econômica[2]. Com o advento da inteligência artificial e aprendizado de máquina, novas estratégias foram criadas para alcançar os objetivos financeiros, utilizando novos conceitos desenvolvidos por essas áreas.

1.2 Utilização de Aprendizado de Máquina em Finanças

O aprendizado de máquina é uma área da ciência da computação que estuda o reconhecimento de padrões e a teoria de aprendizado da inteligência artificial. Por essa razão, esta área está conectada com a estatística computacional, possuindo fortes laços com otimização matemática e reconhecimento de padrões. Alguns exemplos de aplicações incluem reconhecimento de escrita, fala, visão computacional, processamento de linguagem natural e, claro, finanças.

(ZEMKE, 2002) cita os desafios ao se trabalhar com dados financeiros no contexto de aprendizagem de máquina e outros métodos estatísticos. Segundo ZEMKE, o mercado financeiro é um um sistema complexo e não linear, ruidoso, com várias sutilezas

[1] William Feather é um jornalista e escritor norte-americano.
[2] Utilidade econômica é uma medida de satisfação relativa de um agente da economia

e interações dificilmente compreendidas por seres humanos.

Alguns autores utilizaram redes neurais artificiais com diferentes arquiteturas para modelar séries financeiras, como (PANELLA; LIPARULO; PROIETTI, 2014), que utilizaram redes neurais *fuzzy* para modelar o preço de *commodities* do setor de energia e (CHAND; CHANDRA, 2014), que utilizaram coevolução cooperativa[3] para o treinamento da rede neural.

Outra abordagem utilizada no mercado financeiro é a utilização do aprendizado por reforço (RL, do inglês *Reinforcement Learning*) para criação de estratégias de *trading*. (MOODY; WU, 1997) mostram vantagens da utilização do aprendizado por reforço quando comparada à aprendizagem supervisionada.(NEVMYVAKA; FENG; KEARNS, 2006) utilizam esse paradigma de aprendizado para otimizar a execução de ordens em uma bolsa de valores. Em (MOODY; SAFFELL, 1999), são utilizados dois métodos de RL para operações no mercado, *Q-Learning* e *Recurent Reinforcement Learning*. Similarmente, (CORAZZA; BERTOLUZZO, 2014) também usam *Q-Learning* para negociar ações, já (BRITO; ARAUJO, 2013) utilizam o algoritmo SARSA.

Apesar do relativo sucesso apresentado por esses trabalhos, ainda não há sinais claros do impacto do campo de inteligência artificial (IA) na prática, ao se tratar de investimentos. HALL; KUMAR reportaram, por exemplo, que o índice *Eurekahedge AI Hedge Fund*, que segue 12 fundos que usam aprendizado de máquina como parte de suas principais estratégias, apresentou retorno de cerca de 9% ao ano entre 2011 e 2015. Ainda que estes ganhos tenham sido superiores a média de outros *hedge funds*, eles não superaram o retorno do S&P 500[4] no mesmo período.

Contudo, grandes *players* do mercado, como o *Two Sigma Investments*, *Man Group PLC* e *Citadel* apostam no potencial deste campo. Muitos especialistas acreditam que técnicas de IA terão grandes impactos na área de finanças.

1.3 Definição do Problema

Dessa forma, considerando a importância dada para inovação no setor e do potencial do uso de técnicas de inteligência artificial no contexto de investimentos, este trabalho desenvolve uma estratégia adaptativa de investimento, baseado no paradigma de aprendizado de máquina conhecido como aprendizado por reforço (RL).

O RL será explorado com maior detalhe nos próximos capítulos, mas sua ideia geral é encontrar uma estratégia ótima para um problema de decisões em sequência temporal, interagindo diretamente com um ambiente e aprendendo por meio de tentativa e erro.

[3] Algoritmos genéticos onde populações evoluem simultaneamente, tanto para cooperarem entre si como para competirem.

[4] Índice composto pelas ações das 500 maiores empresas americanas

O agente desenvolvido nesse trabalho é baseado no algoritmo *Q-Learning*. Por meio dessa técnica, ele adapta sua estratégia a medida que interage com o ambiente. Seu objetivo é alcançar o máximo de retorno possível, ao final do período de teste. Como método de comparação de desempenho, considera-se o retorno obtido pela ação que está sendo operada no período de tempo considerado.

1.4 Divisão do Texto

Esse trabalho está dividido em seis capítulos. O capítulo dois apresenta a teoria financeira básica para familiarizar o leitor com o contexto do mercado financeiro. O capítulo três apresenta brevemente outros paradigmas de aprendizagem por máquina e aborda com mais detalhes a teoria de aprendizado por reforço. No capítulo quatro, as informações sobre a modelagem do problema são descritas. O capítulo cinco apresenta os resultados obtidos nas simulações. Por fim, o capítulo seis apresenta as conclusões e possíveis extensões desse trabalho.

2 O Mercado de Ações

2.1 Contextualização

O mercado de ações é o ambiente público em que são negociados títulos financeiros, como ações de sociedades de capital aberto e fechado, opções de ações e fundos imobiliários.

As transações são realizadas por intermédio de bolsas de valores. No Brasil, a única bolsa de valores de mercadorias e futuros em operação atualmente é a Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&F BOVESPA).

Esse capítulo apresentará os elementos básicos do mercado financeiro, mais especificamente no que tange a negociação de ações na bolsa de valores. Além disso, serão elucidadas as hipóteses desenvolvidas sobre esses elementos, bem como a metodologia utilizada para a análise de preços.

2.2 Ações

Ações, também chamadas de "papéis", são parcelas do capital social de uma companhia ou sociedade anônima. Elas são títulos patrimoniais, e concedem aos seus titulares (acionistas) os direitos e deveres de um sócio. Dependendo da empresa, o acionista pode receber parte do lucro da mesma por meio de dividendos e juros sobre o capital investido.

Para as empresas, as ações tem o propósito de captar recursos diretamente do público investidor (GRANVILLE, 1963). Isso é feito na forma de abertura de capital, conhecida como Oferta Pública Inicial (IPO, do inglês *Initial Public Offering*). Nesse processo, a empresa vende seus papéis na bolsa de valores pela primeira vez, angariando capital referente ao total de ações vendidas.

A negociação de ações é realizada eletronicamente durante dias úteis e sofre variações diariamente. Se o número de compradores é maior que o de vendedores, o preço da ação tende a subir; caso contrário, o preço tende a cair.

Existem duas principais formas de se operar no mercado de ações: por meio de valor ou por meio de preço.

- **Operar por Valor** : Nessa ótica, a entrada[1] no investimentoé determinada sobretudo pela administração, desempenho ou potencial da empresa. E sua saída[2],

[1] momento onde o investidor realiza a compra
[2] momento onde o investidor realiza a venda

quando os critérios utilizados para a escolha da empresa não estiverem mais sendo satisfeitos.

- **Operar por Preço** : Neste caso, o investimento é focado nas flutuações de preços, inerentes no mercado de ações, no curto e médio prazo. A liquidez (facilidade de se comprar ou vender o ativo) e volatilidade (dispersão dos retornos do ativo) são de extrema importância nesse tipo de metodologia.

Na primeira modalidade, o **valor** percebido da empresa no mercado (presente ou futuro) é o principal fator de análise, sendo o preço atual da ação um fator secundário. Na segunda modalidade, o **preço** atual do papel é o principal fator a ser considerado para realização das decisões.

Nesse trabalho, opera-se por preço. Em outras palavras, o agente inteligente desenvolvido utilizará as flutuações do preço do ativo como base para realizar operações de venda e compra.

2.3 A Dinâmica do Mercado

O lançamento de um IPO representa uma negociação direta entre a companhia e os investidores. Em razão disso, essas transações são realizadas no chamado "mercado primário". Já nas operações realizadas nas bolsas de valores, em que um acionista pode vender suas ações para um potencial comprador, a empresa não participa da negociação. Por esse motivo, o mercado de ações é conhecido como "mercado secundário".

Desse modo, as ações na bolsa de valores são compradas de acionistas que as desejam vender, sem qualquer intermédio da empresa. Nesse cenário, os preços refletem a dinâmica de oferta e demanda. Se os agentes financeiros que fazem parte do mercado, como pequenos investidores, administradores de fundos de investimentos, donos de grandes fortunas, acreditam que a ação irá valorizar, a tendência do preço é aumentar. Isso ocorre porque mais agentes estarão dispostos a comprar e a preços cada vez maiores, já que o preço atual do título não reflete o seu potencial valor futuro (ELDER, 2002). Por outro lado, em um cenário mais pessimista , no qual acredita-se que preço da ação desvalorizará, um número maior de agentes estará disposto a vender a ação por preços cada vez menores.

Um investidor que deseja comprar ações tem duas opções na hora da compra: definir o valor desejado de compra, conhecido como *bid*, ou comprar a preço de mercado, o menor valor de venda oferecido. Para a venda de ações, as alternativas são similares: ou o valor desejado de venda é definido pelo acionista (*ask*), ou a ação é vendida pelo maior valor de compra oferecido. Quando o preço para compra e o preço para venda coincidem, a ordem[3] é executada. O preço efetivo da ação, aquele que é divulgado em relatórios e

[3] Ordem é a informação enviada para a corretora de venda ou de compra

corretoras, corresponde à última ordem executada de um determinado dia. É importante salientar que cada forma de comprar ou vender possui vantagens e desvantagens (GRANVILLE, 1963). As ordens a preço de mercado são executadas quase que instantaneamente, mas não há controle preciso no preço que a ação é negociada. Já quando é feito um *ask* ou *bid*, o agente do mercado tem controle sobre o preço de negociação, mas não há garantia que a ordem será executada.

Considere a situação ilustrada na Tabela 1, em que são apresentadas ordens pendentes de um mercado hipotético.

Preço de Compra (R$)	**Quantidade**	**Preço de Venda (R$)**	**Quantidade**
45,70	100	45,70	100
45,68	150	45,74	200
45,65	200	45,85	200
45,60	100	45,85	150
45,50	150	45,90	100

Tabela 1 – Exemplo de Ordens em um Mercado

Nesse caso, a ordem de compra de 100 ações ao preço de R$ 45.70 será executada pois há uma ordem de venda ao mesmo preço. Caso a ordem de compra a R$ 45.70 fosse de 200 ações, como há apenas 100 ações disponíveis para venda, somente 100 ações seriam compradas a esse preço e uma ordem das 100 ações restantes continuaria pendente.

Suponha agora uma ordem de compra a preço de mercado de 200 ações. As 100 primeiras seriam compradas a R$ 45,70 e o restante a R$ 45,74, resultando em um preço médio de R$ 45,72.

Ao se operar por preço, deseja-se comprar um papel por um preço e, posteriormente, vendê-lo por um preço maior. Isto é, aproveitar os movimentos de alta no mercado. Essa ideia é popularmente conhecida como "comprar barato e vender caro"(ELDER, 2002). Todavia, é possível também obter lucro em movimentos de queda no mercado por meio de prática financeira conhecida como venda a descoberto ou *short*. Nessa operação, o agente financeiro realiza a venda de um papel sem possuir o mesmo, por meio de aluguel, comprando-o de volta no encerramento da posição. Se o preço no momento da venda (início da posição) for maior que o preço na compra (fim da posição), o agente financeiro lucra a diferença de preços. Caso contrário, a posição é encerrada com prejuízo.

Tome como exemplo um *short* executado para 100 papéis que custavam R$ 100.00. O agente que executou essa ordem receberá o valor total de $100 * R\$100,00 = R\$10.000,00$. Se o preço da ação na hora de encerrar a posição estiver R$ 80,00, o valor total de recompra será $100 * R\$80,00 = R\$8.000,00$. Ou seja, a posição foi encerrada com R$ 2,000.00 de lucro.

2.4 Hipóteses do *Random Walk* e do Mercado Eficiente

A capacidade de prever o movimento das ações e de encontrar padrões nas variações de preços são debatidas frequentemente na teoria financeira. As duas principais hipóteses que refutam essa capacidade são conhecidas como Hipótese do *Random Walk* e Hipótese do Mercado Eficiente (HME).

A Hipótese do *Random Walk* foi desenvolvida pelo matemático francês Louis Bachelier na sua tese de doutorado (BACHELIER, 1900). Nesse trabalho, o autor afirma que o movimento de preço de uma ação pode ser modelado por um processo estocástico conhecido como *Random Walk*. Esse processo pode ser interpretado como uma forma de descrever um caminho de passos aleatórios. Apesar da popularidade dessa teoria, vários autores já refutaram essa hipótese. Em (LO; MACKINLAY, 1988), a hipótese é rejeitada com base em testes estatísticos de autocorrelação. Além desse trabalho, outras obras como (FRENCH, 1988) e (COHEN et al., 1983) obtiveram indicativos de previsibilidade no preço de ações.

Já a HME é uma teoria econômica que afirma que o preço de um ativo financeiro é reflexo do nível de informação disponível sobre o mesmo. Dessa forma, seria impossível obter um desempenho superior ao do mercado, visto que os preços só reagiriam a novas informações e, por conseguinte, dados históricos não teriam utilidade para prever as movimentações no preço do ativo. Em outras palavras, o mercado é **eficiente** ao absorver novas informações. Essa hipótese foi desenvolvida por Eugene Fama no seu trabalho (FAMA, 1970). O autor afirma que os ativos são negociados a "preço justo"e que seria impossível obter retorno em excesso, ou seja, comprar ações subvalorizadas ou vender ações supervalorizadas.

A HME requer que os agentes financeiros tenham expectativas racionais. Isso quer dizer que, na média, a população tem expectativas corretas de acordo com o nível de informação disponível, mesmo que alguns agente individuais reajam exageradamente. Em outras palavras, o mercado sempre está certo.

Existem três hipóteses para definir a eficiência dos mercados.

- **Eficiência Fraca** - Os preços não podem ser previstos analisando dados históricos. Não existem padrões ou correlação dos preços passados com os preços futuros e são determinados exclusivamente por informações que não estão contidas na série financeira. Logo, os preços seguem um *Random Walk*.

- **Eficiência Semiforte** - Os preços dos papéis negociados são ajustados por novas informações divulgadas publicamente, de maneira quase instantânea. Desse modo, é impossível obter retornos em excesso realizando operações com base nessas informações.

- **Eficiência Forte** - Nessa forma, o preço da ação reflete todas as informações, sejam públicas ou privadas, e não é possível obter retorno em excesso.

A ineficiência do mercado já foi tratada em trabalhos como (GROSSMAN; STIGLITZ, 1980) e (BLACK, 1986). Além disso, análises empíricas e testes experimentais têm fornecido evidências de que, em muitas situações, as decisões financeiras desviam-se bastante daquelas implicadas pelo tipo de racionalidade na qual a HME se sustenta (ALDRIGHI; MILANEZ, 2005). Bolhas financeiras, volatilidade excessiva dos preços dos ativos e aquisições de empresas a termos desfavoráveis são exemplos dessas decisões "irracionais".

O mercado de ações é marcado pelas interações entre inúmeros agentes financeiros participantes, que nada mais são do que seres humanos. Por consequência, há uma vertente em economia denominada Finanças Comportamental (FC), que tem desenvolvido conceitos e teorias baseadas sobretudo em limitações intrínsecas dos indivíduos, que os impedem de realizar decisões estritamente racionais.

Por fim, outra teoria que vai contra os preceitos da HME é conhecida como Hipótese dos Mercados Adaptativos (HMA) (LO, 2004). Essa hipótese baseia-se na interpretação evolutiva das interações econômicas aliada com noções de neurociência. A HMA pode ser vista como uma nova versão da HME. Os preços refletem tanta informação quanto a combinação das condições do ambiente (mercado) e do número e natureza das espécies (agentes financeiros) da economia.

2.5 Formas de Análise

Quanto a compra e venda de um ativo, existem duas principais metodologias nas quais os investidores se baseiam para a tomada de decisão: análise técnica e análise fundamentalista. Cada abordagem utiliza técnicas diferentes para fazer a avaliação de uma ação. Apesar disso, elas não são mutualmente excludentes e inúmeros investidores utilizam as duas abordagens para a escolha de papéis. Muitas vezes, a Análise Fundamentalista é usada para definir em qual empresa ou setor investir, e a técnica é utilizada para escolher o momento de entrada e saída no mercado.

2.5.1 Análise Fundamentalista

A análise fundamentalista baseia-se na avaliação das características de um empresa com o intuito de estimar o seu valor (GRAHAM, 1965). São analisados dados do setor da economia do qual a empresa se enquadra, sua situação fiscal, seu fluxo de caixa, indicadores regionais e globais que possam influenciar a lucratividade da empresa, entre

outros. O objetivo principal da análise fundamentalista é encontrar o valor intrínseco de um papel.

O valor intrínseco pode ser visto como uma espécie de "valor real" de uma ação, espera-se que o preço da ação se iguale ao valor intrínseco com o passar do tempo. Esse valor é utilizado pelos investidores para ajudar na tomada de decisões, comparando-o com o preço atual da ação. Um valor intrínseco maior do que o valor atual da ação é considerado um indicativo para compra do ativo. Por outro lado, se o valor intrínseco for menor que o valor atual, recomenda-se vender o ativo ou realizar um *short*.

É importante ressaltar que o conceito de valor intrínseco se opõe ao que a HME afirma. Segundo essa abordagem, o preço atual da ação é reflexo de todas as informações disponíveis, não havendo espaço para um possível valor intrínseco ser atingido. Além disso, muitos fatores investigados por essa metodologia são qualitativos, tais como: o atual presidente da empresa, o nível de organização e a imagem da companhia na mídia. Dessa forma, é difícil fazer uma avaliação estatística da real eficiência desse tipo de análise.

2.5.2 Análise Técnica

A outra metodologia extremamente utilizada no mercado financeiro é a Análise Técnica, que considera os dados históricos de uma ação para tomada de decisões. O principal pressuposto dessa abordagem é que os dados históricos da ação são suficientes para justificar a realização de operações financeiras. Em outras palavras, acredita-se que toda a informação da empresa necessária está presente nos dados passados (DAVIDSON, 2000)(ELDER, 2002) . Foram desenvolvidos indicadores que visam utilizar tais informações para embasar a tomada de decisões, que serão apresentados na próxima seção.

A Análise Técnica defende que os preços de um ativo seguem tendências, ou seja, uma ação com tendência de alta geralmente apresenta um movimento de crescimento nos preços, o contrário aplica-se para um ativo com tendência de baixa, conforme ilustrado na Fig. 1.

Outra ideia defendida pela análise técnica é que a história se repete. Em outras palavras, o comportamento dos investidores tende a se repetir de tempos em tempos. Portanto, existem padrões que podem ser identificados e explorados pelos analistas técnicos.

Um desses padrões é intitulado como suporte e resistência, que caracteriza-se como a dificuldade que os preços têm para se mover acima de uma resistência ou abaixo de um suporte, vide Fig. 2.

As linhas de suporte e resistência são vistas como fatores de psicologia do mercado e representam os níveis em que muitos investidores estão dispostos a comprar ou vender as ações.

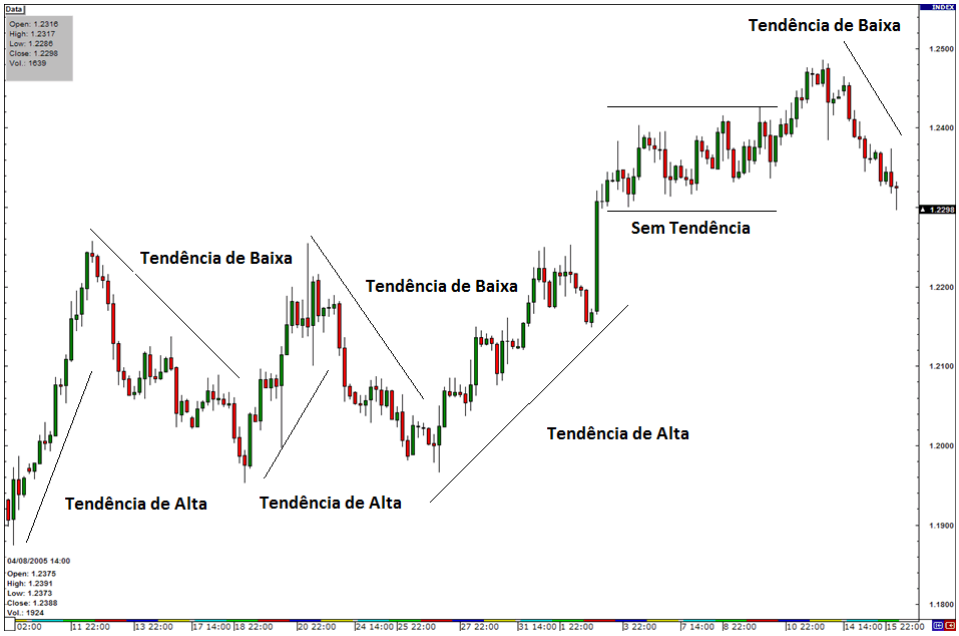

Figura 1 – Movimentos de um Ativo

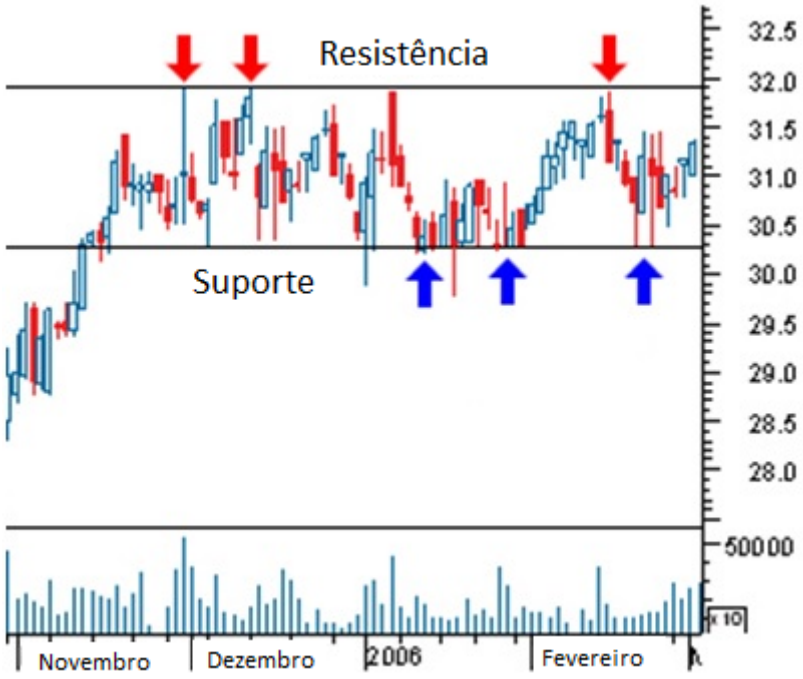

Figura 2 – Resistência e Suporte

Segundo os analistas técnicos, uma vez que um nível de resistência ou suporte é quebrado, o papel sofre uma reversão, ou seja, se o preço cai abaixo do nível de suporte, esse nível será a nova resistência. Caso o preço suba acima do nível de resistência, ele se tornará o novo suporte. Observa-se esse fenômeno na Fig. 3.

Uma última hipótese central é a de Retorno à Média. Essa ideia é similar ao do Valor Intrínseco, utilizada na Análise Fundamentalista. Ou seja, apesar das possí-

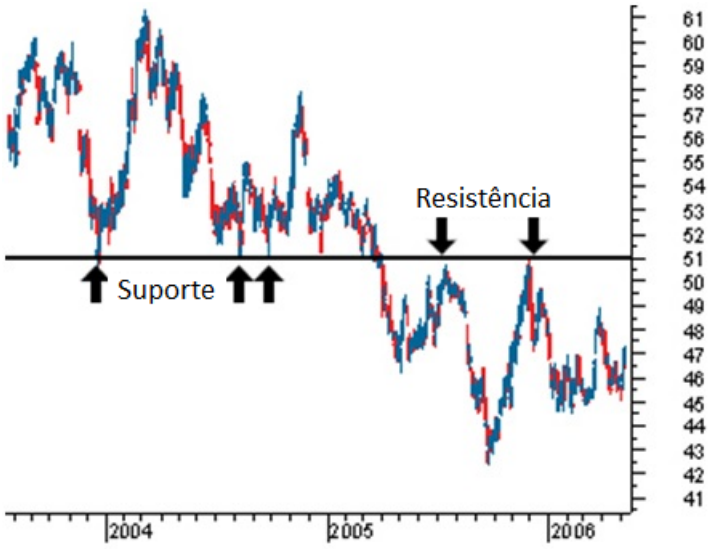

Figura 3 – Inversão de um papel

veis variações que ocorrem no preço da ação, ele tende a retornar ao seu Valor Médio. Tais variações são exploradas por alguns indicadores técnicos para obtenção de operações lucrativas.

É importante ressaltar que, apesar da popularidade da Análise Técnica e da sua utilização por diversos investidores, não existe comprovação estatística da sua real eficiência. Isso se deve ao fato de que muitas vezes as decisões, segundo uma visão técnica, são tomadas com base em tendências e formatos de gráficos, fatores muito subjetivos para serem avaliados rigorosamente.

Contudo, este trabalho é embasado nesse tipo de análise, utilizando indicadores como parâmetros de entrada do programa que realizará a tomada de decisões.

2.6 Indicadores

Indicadores são medidas estatísticas, utilizados na análise técnica, que foram desenvolvidos com o intuito de fornecer mais informações do que simplesmente o preço do ativo ou o volume negociado de ações em determinado período. Os indicadores utilizados nesse trabalho serão apresentados nas seções subsequentes.

2.6.1 Retorno Diário

O retorno diário fornece informação sobre o desempenho da ação no dia n em relação ao dia $n-1$, calculado da seguinte forma:

$$r_n = \frac{P_n}{P_{n-1}} - 1, \tag{2.1}$$

em que P_n é o preço de fechamento do ativo no dia n. Esse indicador pode ser utilizado para observar tendências de alta ou baixa. Se uma ação apresentar uma sequência de retornos diários positivos, isso pode indicar uma tendência de valorização do preço.

Ao contrário de muitos indicadores que utilizam dados históricos de um período de tempo, o retorno diário depende somente do preço no dia atual e no anterior. Por esse motivo, ele sinaliza com rapidez uma mudança repentina de preços.

2.6.2 Média Móvel

O retorno diário, apesar de ter sua devida importância na análise técnica, pode fornecer uma visão muito imediatista sobre o desempenho de determinado ativo.

A identificação de movimentos, em médio ou em longo prazo, pode ser feita por meio de um indicador conhecido como média móvel. Apesar dessa técnica não ser utilizada para prever a direção do preço em curto prazo, ela "suaviza" os preços e define a sua tendência de movimento. Diante disso, esse indicador é classificado como indicador de tendência.

A média móvel simples, ou *Simple Moving Average*, (SMA) de N períodos pode ser calculada por:

$$SMA_N = \frac{1}{N} \sum_{n=1}^{N} P_n. \tag{2.2}$$

Uma possível interpretação da média móvel pode ser feita comparando seu valor com o preço atual do ativo. Se o preço do ativo estiver maior que a média móvel de um período razoavelmente longo, por exemplo 20 ou 100 dias, isso pode indicar uma tendência de alta e o mesmo vale para um preço abaixo da média móvel. Porém, é importante notar que uma variação brusca no preço da ação causada por algum distúrbio no mercado pode causar instabilidade nesse indicador. Isso ocorre a partir do momento que esse dia "anormal" deixa de ser considerado no cálculo da média móvel.

Para lidar com esse fenômeno, foi criada a chamada média móvel exponencial ou *exponential moving average* (EMA) (CHAO-WEN; JR, 1999). Esse indicador pondera com peso maior as mudanças de preços que ocorreram recentemente. Pode-se calcular a EMA da seguinte forma:

$$EMA_N = (P_n - EMA_{n-1}) * \frac{2}{N+1} + EMA_{n-1}. \tag{2.3}$$

Para o cálculo inicial da EMA_N, é utilizada a SMA_N do ativo em questão. Após o cálculo inicial, os valores são calculados de forma recursiva. O fator $\frac{2}{N+1}$ é conhecido como fator multiplicativo e seu objetivo é atribuir um peso diferente ao preço mais recente.

2.6.3 Oscilador Estocástico

O Oscilador Estocástico é um indicador técnico de momento baseado em níveis de suporte e resistência (MURPHY, 1999) . Essa classe de indicadores é utilizada para medir as taxas de variação no preço de um papel. O oscilador estocástico tem dois fatores, $\%K$ e $\%D$ que são calculados por:

$$\%K_n = \frac{P_n - L_{14}}{H_{14} - L_{14}} * 100, \tag{2.4}$$

$$\%D_n = \frac{\%K_n + \%K_{n-1} + \%K_{n-2}}{3}; \tag{2.5}$$

em que P_n é o preço de fechamento do ativo no dia n, L_{14} e H_{14} são, respectivamente, o menor e o maior valor atingido pelo ativo nos últimos 14 períodos.

A variável $\%K$ mede o nível de proximidade do preço atual em relação à variação dos preços máximo e mínimo no período analisado. Seu valor varia de 0 a 100: um valor superior a 80 indica que o ativo está próximo do seu máximo no período e um valor inferior a 20 indica que ele está próximo de seu mínimo. A variável $\%D$ é plotada ao lado de $\%K$ e serve de "gatilho" para o investidor. Ou seja, uma operação compra ou venda é sinalizada quando $\%D$ e $\%K$ se tocam.

A Fig. 4 mostra o oscilador estocástico com $\%K$ e $\%D$ sendo representados na parte inferior.

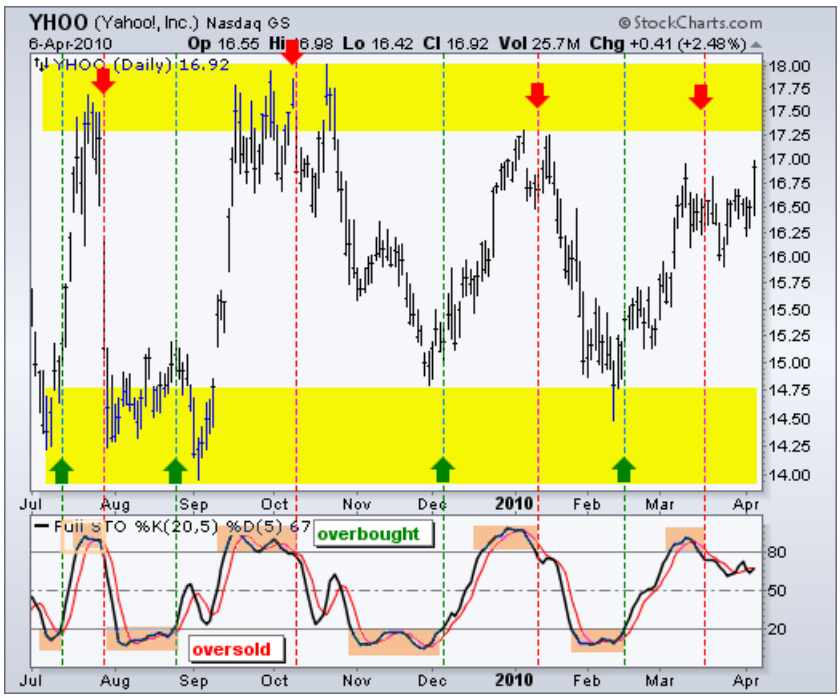

Figura 4 – Oscilador Estocástico, retirado de (STOCKCHART, 2017)

2.6.4 Bandas de Bollinger

Na década de 80, o analista John Bollinger desenvolveu um novo indicador, que foi denominado de bandas de Bollinger. Essa técnica consiste na criação de uma região delimitada por duas curvas, intituladas bandas superior e inferior. Elas são formadas utilizando a média móvel dos preços de um ativo acrescida de K vezes o valor do desvio padrão do ativo para formar a banda superior (BS) e decrescida desse mesmo valor para a banda inferior (BI), como pode ser visto nas equações 2.6 e 2.7.

$$BS = SMA_N + K * \sigma_N, \tag{2.6}$$

$$BI = SMA_N - K * \sigma_N, \tag{2.7}$$

em que SMA é a média móvel e σ_N é o desvio padrão do preço da ação no período de N dias. Valores típicos para N e K são 20 e 2, respectivamente.

Existem diversas interpretações sobre a forma de utilizar as bandas de Bollinger. Uma das possíveis estratégias é considerar como indicativo de compra ou venda o momento no qual o preço da ação toca em alguma das bandas. A justificativa baseia-se na ideia de Retorno à Média. O distanciamento em mais de dois desvios padrões do preço médio é considerado um evento incomum. Portanto, há a tendência do preço voltar ao seu Valor Médio.

Figura 5 – Bandas de Bollinger

2.6.5 Média Móvel Convergente e Divergente

O último indicador a ser apresentado nesse trabalho é a Média Móvel Convergente e Divergente, popularmente conhecido como MACD (*Moving Average Convergence Di-*

vergence). Um dos aspectos mais interessantes do MACD é a utilização de indicadores de tendência (média móveis) para indicar momento.

O MACD, assim como o oscilador estocástico, tem dois componentes: a "Linha MACD"(LM), formada pela diferença das médias móveis exponenciais de 12 e 26 períodos, e a "Linha Sinal"(LS), formada pela média móvel exponencial de 9 períodos da LM. Os intervalos de tempo considerados para o cálculo podem variar de acordo com a estratégia do operador. Os valores descritos foram os definidos por seu criador (APPEL, 1985).

$$LM = EMA_{12} - EMA_{26}, \tag{2.8}$$

$$LS = EMA_9(LM). \tag{2.9}$$

Como o nome do indicador sugere, ele baseia-se na convergência e na divergência das médias móveis. A primeira ocorre para pequenos amplitudes de LM, que indicam que não há diferença relevante entre a EMA_{12} e EMA_{26}, já a divergência é indicada por grandes amplitudes de LM. Valores positivos de LM sugerem um aumento no momento de alta. Similarmente, valores negativos sugerem um aumento no momento de baixa da ação.

Apesar do valor de LM já conter informações relevantes sobre a ação, a utilização da LS é o que o torna completo. Um sinal de compra é definido quando a LM cruza a LS em um movimento crescente. Quando o sentido da interceptação é um movimento decrescente, define-se um sinal de venda.

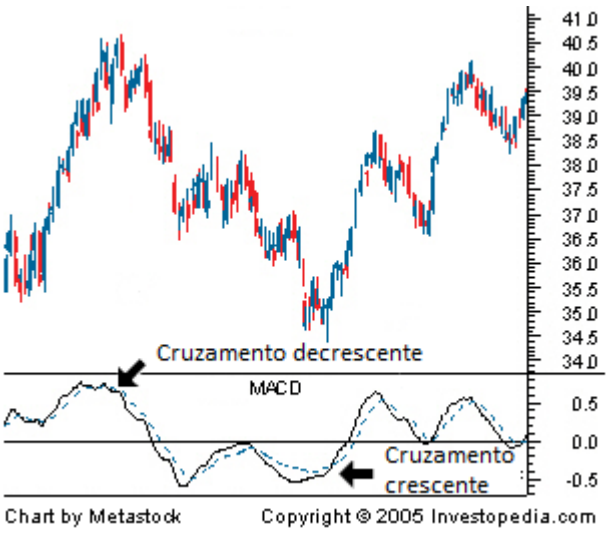

Figura 6 – Média Móvel Convergente e Divergente, adaptado de (INVESTOPEDIA, 2017)

A Fig.6 exemplifica esses movimentos. No primeiro momento destacado, a LM, representada pela linha preta, cruza a LS para baixo, indicando uma possível venda. Em

um segundo momento, ocorre o movimento contrário, com a LM cruzando a LS para cima, sinalizando a compra.

3 Aprendizado por Reforço

3.1 Contexto

O Aprendizado por Reforço ou *Reinforcement Learning* (RL) é um dos paradigmas de Aprendizado de Máquina. Diferente da forma supervisionada de aprendizado, o RL não conta com um "supervisor". Existe um agente inteligente que toma ações e observa o sinal do ambiente denominado sinal de recompensa. De acordo com esse sinal, o agente reorganiza suas decisões, aprendendo por meio de tentativa e erro.

3.2 Outros Paradigmas de Aprendizado de Máquina

3.2.1 Aprendizado Supervisionado

O paradigma de aprendizado de máquina mais estudado atualmente é o Aprendizado Supervisionado. Nessa abordagem, a aprendizagem é realizada por meio de dados de treinamento já classificados que são fornecidos por um "supervisor externo"(SUTTON; BARTO, 1998).

Cada exemplo é uma descrição de uma situação em que a saída correta já é conhecida. Considere um problema de classificação de dígitos, a imagem de entrada no treinamento é acompanhada da saída desejada, ou seja, o dígito correspondente. Ao fim do treinamento, espera-se que o sistema consiga generalizar sua saída para dados de entrada não vistos anteriormente.

Entretanto, em problemas iterativos, torna-se praticamente inviável a obtenção de exemplos de comportamentos adequados e, ao mesmo tempo, que representem as situações possíveis que um agente pode encontrar.

3.2.2 Aprendizado Não-supervisionado

Outro paradigma existente no aprendizado de máquina é o Aprendizado Não-supervisionado. Diferente do caso supervisionado, não há um professor que indique a saída esperada para os exemplos de treinamento. O objetivo é, na maioria dos casos, encontrar padrões e outras estruturas ocultas contidas nos dados.

Segundo (SUTTON; BARTO, 1998), apesar de também não contar com exemplos de comportamentos corretos, o Aprendizado por Reforço diferencia-se do não-supervisionado pelo seu principal objetivo. Em vez de encontrar padrões ou estruturas ocultas, busca-se maximizar a recompensa total obtida.

3.3 Elementos do Aprendizado por Reforço

O Aprendizado por Reforço tem duas características principais. A primeira delas é que o agente aprende por meio de tentativa e erro, ou seja, de acordo com o *feedback* recebido do ambiente. A segunda característica é a noção de recompensa tardia. O agente deve levar em conta recompensas futuras que podem ser recebidas posteriormente ao se realizar determinada ação no presente. Em outras palavras, o agente deve considerar a temporalidade do problema ao tomar uma ação.

O Aprendizado por reforço dá ênfase na aprendizagem pela interação direta do agente com o ambiente em termos de estados, ações e recompensas.

3.3.1 Recompensa

A Recompensa R_t é um sinal de *feedback* escalar. Seu objetivo é indicar ao agente o quão bem ele está agindo no ambiente (SILVER, 2015). A principal função do agente é **maximizar** a recompensa cumulativa recebida por meio da escolha correta de **ações**.

As ações tomadas podem ter recompensas tardias e consequências no cenário futuro enfrentado pelo agente. Por esse motivo, algumas vezes recompensas em curto prazo são sacrificadas em prol de recompensas em longo prazo melhores.

3.3.2 Agente e Ambiente

No Aprendizado por Reforço, há a interação direta entre o **Agente** e o **Ambiente**. Em cada instante de tempo t, o agente executa uma ação A_t, recebe uma observação O_t e uma recompensa R_t. Já o ambiente, recebe uma ação do agente A_t, emite uma observação O_{t+1} e uma recompensa R_{t+1}. O tempo é então incrementado e o processo continua. Essa relação entre agente e ambiente pode ser observada na Fig.7.

Outro conceito indispensável é o de **estado** S_t. Ele contém as informações que são utilizadas para determinar o que acontecerá em seguida. O agente toma suas ações de acordo com seu estado, S_t^a. O objetivo de S_t^a é representar da melhor forma possível o estado do ambiente, denominado S_t^e.

Em algumas situações, o agente pode observar diretamente o estado do ambiente. Logo, $O_t = S_t^e = S_t^a$. Esses ambientes são chamados de completamente observáveis.

Em ambientes parcialmente observáveis não é possível obter diretamente S_t^e. Cabe então ao agente construir uma representação S_t^a que seja capaz de fornecer informações suficientes para a tomada de decisões, por meio da observação indireta do ambiente.

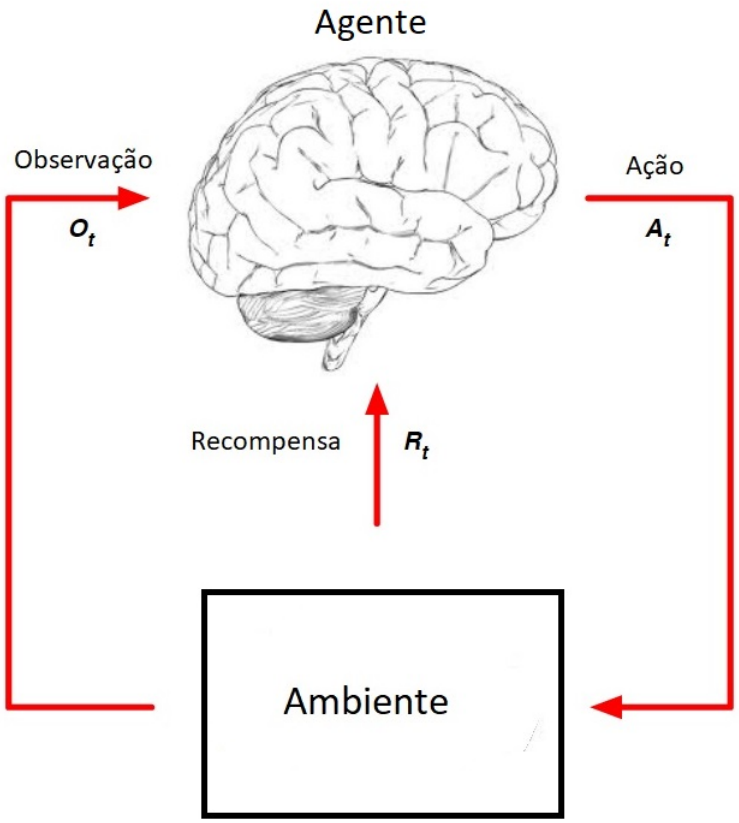

Figura 7 – Representação de um problema de Aprendizado por Reforço

3.4 Processos de Decisão de Markov

Como já descrito anteriormente, o agente no RL utiliza a informação contida nos estados para decidir qual ação tomar. Porém, o estado não contém obrigatoriamente todas as informações do ambiente. Por exemplo, se o agente está participando de um jogo de cartas, não se deve esperar que ele tenha informação de quais cartas que estão viradas para baixo. Idealmente, espera-se que a representação de estados consiga fornecer informação suficiente sobre o problema de forma compacta.

Com isso em mente, pode-se definir o conceito de propriedade de Markov para a representação dos estados. Um estado possui a propriedade de Markov se a resposta do ambiente no tempo $t+1$ só depende da ação e do estado no tempo t. Em outras palavras, o estado atual apresenta todas as informações necessárias para o agente tomar uma ação. Formalmente, temos:

$$P[S_{t+1} = s', R_{t+1} = r \mid S_0, A_0, R_1, ..., S_{t-1}, A_{t-1}, R_t, S_t, A_t] = P[S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a]. \tag{3.1}$$

A propriedade de Markov é importante no aprendizado por reforço porque assume-

se que as decisões e os valores são funções apenas do estado atual. Ou seja, a representação do estado precisa ser informativa. Um problema de RL que satisfaz a propriedade de Markov é chamado de processo de decisão de markov (MDP, do inglês *Markov Decision Process*).

Segundo (SILVER, 2015), um MDP é definido formalmente por uma quádrupla $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, onde:

- $\mathcal{S}$ é o espaço finito de estados que contém todos os estados possíveis $(s_1, s_2, ...s_n)$
- $\mathcal{A}$ é o espaço finito de ações que contém todas as ações possíveis $(a_1, a_2, ...a_n)$
- $\mathcal{P}$ é a matriz de transição de estados, ela representa a probabilidade de sair do estado s para o estado s' após tomar a ação a, $\mathcal{P}^a_{ss'} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$.
- $\mathcal{R}$ é a função de recompensa, ela define a punição ou bonificação do agente ao tomar a ação a no estado s, $\mathcal{R}(s, a) = \mathbf{E}[R_{t+1} \mid S_t = s, A_t = a]$, onde $\mathbf{E}[.]$ é o valor esperado de uma variável aleatória.

Um MDP é definido pelo seu estado, ação e a dinâmica do ambiente no próximo instante de tempo. Dada qualquer ação a e estado s, a probabilidade de cada par futuro de recompensa-estado s', r possível é dado por:

$$p(s', r \mid s, a) = P[S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a]. \tag{3.2}$$

3.5 Função Valor

Outro elemento fundamental do aprendizado por reforço é a função valor. Essa função utiliza o estado (ou par ação-estado) para estimar o quão bom é estar em determinado estado (ou o quão bom é tomar determinada ação para um dado estado). O conceito de "quão bom" é definido em termos de futuros retornos esperados.

O retorno G_t é definido como a soma descontada das recompensas a partir do instante t (SUTTON; BARTO, 1998).

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{3.3}$$

O termo γ é a taxa de desconto, ele determina o valor presente de recompensas futuras. Um valor de γ próximo de 0 levará o agente a priorizar recompensas mais imediatas. Da mesma forma, o agente valorizará recompensas em longo prazo quando γ é próximo de 1.

A Função Valor é definida de acordo com uma determinada política. A política π pode ser interpretada como o mapeamento de cada estado $s \in \mathcal{S}$ e ação $a \in \mathcal{A}$ com a probabilidade $\pi(a|s)$ do agente tomar a ação a dado o estado s. O **valor** de um estado s seguindo a política π, denotado por v_π, é, simplesmente, o retorno esperado ao seguir a política π estando originalmente no estado s.

$$v_\pi(s) = \mathbf{E}_\pi[G_t \mid S_t = s], \tag{3.4}$$

em que $\mathbf{E}_\pi$ é o valor esperado de uma variável aleatória, dado que o agente seguiu a política π.

Do mesmo modo, define-se a função ação-valor $q_\pi(s, a)$ como o valor de se tomar determinada ação a, em um estado s e, posteriormente, seguir uma política π. Formalmente, temos:

$$q_\pi(s,a) = \mathbf{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbf{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a] = \sum_{k=0}^{\infty} \gamma^k R(s_k, a_k). \tag{3.5}$$

Pode-se separar a Eq.(3.5) em duas partes:

$$\begin{aligned} q_\pi(s,a) &= \mathcal{R}(s,a) + \sum_{k=1}^{\infty} \gamma^k R(s_k, a_k) \\ &= \mathcal{R}(s,a) + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} R(s_k, \pi(s_k)) \\ &= \mathcal{R}(s,a) + \gamma v_\pi(s'). \end{aligned} \tag{3.6}$$

O primeiro componente é a recompensa imediata $\mathcal{R}(s, a)$ de tomar a ação a no estado s. O segundo termo é o retorno esperado ao seguir a política π a partir do estado s'.

A resolução de um problema de RL consiste em encontrar a política ótima π_*, definida como a política que apresenta o maior retorno esperado. Podemos estender essa definição para a função valor e função ação-valor:

$$v_*(s) = \max_\pi \; v_\pi(s), \tag{3.7}$$

$$q_*(s,a) = \max_\pi \; q_\pi(s,a), \tag{3.8}$$

para todo $s \in \mathcal{S}$ e $a \in \mathcal{A}$. Como o par estado-ação (s,a) representa o retorno esperado de se tomar a ação a no estado s e seguir a política ótima posteriormente, pode-se então escrever q_* em função de v_* da seguinte forma:

$$q_*(s,a) = \mathbf{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]. \tag{3.9}$$

3.6 Política

O agente no RL deve realizar uma escolha fundamental no decorrer de seu treinamento, isto é, deve-se decidir entre explorar novas opções (*exploration*) ou escolher uma opção já conhecida (*exploitation*).

Na seção anterior, a política foi definida como o mecanismo que faz o agente escolher uma ação $a \in \mathcal{A}$ em um estado $s \in \mathcal{S}$. Utilizando o valor da função Q, a estimativa de q_π, podemos realizar essa escolha de diversas maneiras, entre elas:

Força Bruta

Testa todas as políticas possíveis e escolhe a que retornar a maior recompensa. Essa técnica é extremamente ineficiente e pode se tornar inviável em problemas de maior escala.

Política Ganaciosa (*Greedy*)

Essa técnica escolhe a ação que possui o maior valor naquele estado, que é equivalente a escolher a ação do $Q(s', a)$ que possuir o maior valor no estado s'. O problema dessa abordagem é a falta de exploração de novas opções (*exploration*). O agente fica limitado ao conhecimento disponível atualmente.

Política ϵ-Ganaciosa (ϵ-*Greedy*)

Uma das políticas mais utilizadas. A ação é escolhida ϵ% das vezes aleatoriamente. Os $(1 - \epsilon)$% restantes são escolhidos de forma gananciosa. Dessa forma, o agente pode desbravar novas opções quando a ação for aleatória e explorar o que já é conhecido de forma gananciosa. Uma forma mais eficiente de implementar essa política é iniciar o valor de ϵ em 1 e ir reduzindo seu valor a medida que o agente for explorando as ações possíveis.

3.7 Aprendizagem por Diferenças Temporais

De acordo com (SUTTON; BARTO, 1998), um dos pontos principais por trás do RL é a ideia de aprendizagem por diferenças temporais ou *temporal-difference learning*

(TD). Existem outros métodos que podem ser aplicados no aprendizado por reforço, como Monte Carlo ou programação dinâmica. Por questões de brevidade, somente o método TD será aprofundado no texto, visto que é a abordagem utilizada pelo algoritmo *Q-Learning*, apresentado na próxima seção.

Os métodos TD conseguem aprender diretamente com base nas experiências, não necessitando de um modelo dinâmico do ambiente. Além disso, eles aprendem de forma *online*, aprendendo em todo instante, e não só no final do episódio[1], como um método *offline* (Monte Carlo). Pode-se observar essa característica na equação a seguir, um exemplo simples do método de Monte Carlo:

$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]. \tag{3.10}$$

O valor de $V(S_t)$ é atualizado com base no valor de G_t e de uma constante α. Porém, o valor de G_t só é conhecido ao final de um episódio. Em contrapartida, o método TD mais simples (TD(0)) é enunciado da seguinte forma:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]. \tag{3.11}$$

Nota-se que só é preciso chegar ao instante $t+1$ para se atualizar o valor de $V(S_t)$. Por essa razão, os métodos TD conseguem aprender mesmo em episódios ainda incompletos. O termo entre colchetes na equação 3.11 pode ser interpretado como uma forma de erro, visto que ele mede a diferença entre $V(S_t)$ e sua estimativa $R_{t+1}+\gamma V(S_{t+1})$. Essa questão será discutida com mais profundidade quando na seção de aproximação de função, localizada no Capítulo 4.

As mesmas derivações feitas para a função valor $V(S_t)$ podem ser aplicadas para a função ação-valor $Q(S_t, A_t)$. Essa é peça fundamental do algoritmo *Q-Learning*.

3.8 *Q-Learning*

Um dos primeiros grandes avanços do aprendizado por reforço foi o desenvolvimento do algoritmo *Q-Learning* por (WATKINS, 1989), definido como:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a') - Q(S_t, A_t)], \tag{3.12}$$

no qual $\alpha \in (0, 1]$ é chamado taxa de aprendizagem e $a' \in \mathcal{A}$ é a ação que maximiza o valor de $Q(S_{t+1}, a)$. O termo entre chaves é o erro de estimação no tempo t, também

[1] Um episódio é caracterizado como o período de treinamento no qual o agente atua, começando pelo estado inicial e finalizando no estado terminal

conhecido como erro TD. Nesse caso, a função Q aprendida se aproxima da função ação-valor ótima q_*, independente da política utilizada. O único requisito para convergência é que os pares ação-estado (a, s) continuem sendo atualizados, esse fato é provado em (WATKINS; DAYAN, 1992).

O *Q-Learning* é considerado um algoritmo *off-policy*. Isso decorre do fato do valor do $Q(S_t, A_t)$ ser atualizado com base no valor $Q(S_{t+1}, a')$, onde a' é a ação com o maior valor de Q no estado S_{t+1}. Ou seja, independente da política utilizada para escolher a ação, Q é atualizado com base em uma política gananciosa.

Outro algoritmo utilizado em problemas de RL é conhecido como SARSA. Esse é bastante similar ao algoritmo do *Q-Learning*, a diferença está na atualização do valor de $Q(S_t, A_t)$. Em vez de usar a política gananciosa para atualizar o valor de $Q(S_t, A_t)$, o algoritmo usa o valor de $Q(S_t, a'')$, onde a'' é a ação escolhida pela atual política do agente. Por essa razão, o SARSA é considerado um algoritmo *on-policy*. Esse fato pode ser visto com mais clareza na equação Eq.(3.13).

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma\ Q(S_{t+1}, a'') - Q(S_t, A_t)]. \tag{3.13}$$

Para problemas de RL mais simples, com poucos estados ou ações possíveis, os valores de $Q(s, a)$ podem ser armazenados em uma tabela (*lookup table*) ou matriz, como pode ser observado na Tabela 2.

$Q(s, a)$	**Estado 1**	**Estado 2**	**Estado 3**	**Estado 4**
Ação 1	10	7	12	25
Ação 2	15	4	-13	32
Ação 3	-20	45	20	20
Ação 4	12	-11	10	-10

Tabela 2 – Exemplo de Valores de Q

Contudo, em casos com muitos estados discretos ou com um espaço de estados contínuo, é inviável realizar essa conferência para cada instante de tempo t. Opta-se então por usar uma função $\mathcal{F}(s, a)$ que aproxime o valor de Q. Essa aproximação será detalhada no capítulo seguinte.

4 Desenvolvimento do Agente Inteligente

4.1 Considerações Iniciais

Os Capítulos 3 e 4 apresentaram conceitos básicos do mercado financeiro e do aprendizado por reforço. Esse capítulo tem como objetivo apresentar a forma na qual essas duas áreas foram unidas para desenvolver o agente inteligente.

As operações que ocorrem na bolsa de valores nada mais são do que relações de compra e venda, realizadas a todo instante, desde o início até o final do pregão.

Devido a essas características, algumas simplificações e considerações foram feitas de modo a tornar a modelagem do problema viável. Apesar disso, não há prejuízo significativo na interpretação dos resultados.

Os dados financeiros utilizados foram obtidos em intervalos de 1 em 1 dia. Ou seja, o instante "$t + 1$" representa um dia após t. Consequentemente, os preços de abertura e fechamento refletem a variação em um período de operação de um dia na bolsa de valores. Considera-se que o agente realiza todas as suas operações no fechamento do pregão, em que o preço vigente é dado pelo preço de fechamento (*closing price*). Ou seja, o agente não realiza ações no meio do pregão ou no seu início.

Outra simplificação feita diz respeito aos custos de transação. Foram desconsiderados quaisquer custos de operação como taxas de corretagem, custódia e outros valores que podem fazer parte de uma operação real na bolsa. Essa simplificação, apesar de não refletir o cenário real, não provoca consequências práticas expressivas. Para um montante razoavelmente grande de capital, esses custos não tem impacto significativo no rendimento da operação.

A última consideração a ser feita é quanto ao número de ativos operados ao mesmo tempo. Por questões de simplicidade, somente uma ação foi considerada por vez. Cada episódio consistia no agente acompanhando as variações de preço do ativo em questão, aprendendo durante o tempo estabelecido e, posteriormente, operando esse mesmo ativo no período de teste.

4.2 Codificação das Ações

O sistema desenvolvido neste trabalho pode realizar três decisões distintas, são elas:

- **Comprar**: adquire o máximo de unidades possível do ativo. Corresponde à ação

"1".

- **Manter**: mantém a posição atual. Corresponde à ação "0".
- **Vender / *Short***: vende todas as unidades na carteira. Se o agente estiver fora do mercado, entra em posição de *short*. Corresponde à ação -1".

É importante ressaltar a definição de "posição". Se o agente está fora do mercado (sem nenhum capital investido), ele encontra-se em uma posição neutra. Caso opte por comprar ações, ele passa a estar em uma posição comprada. Similarmente, fazer um *short* o coloca em uma posição vendedora.

Caso o agente escolha comprar em um instante de tempo no qual ele já está em uma posição comprada, a posição é mantida. O mesmo vale para a venda em uma posição vendida. A ação "manter"preserva a posição anterior do agente. Essa interação é ilustrada na Fig. 8.

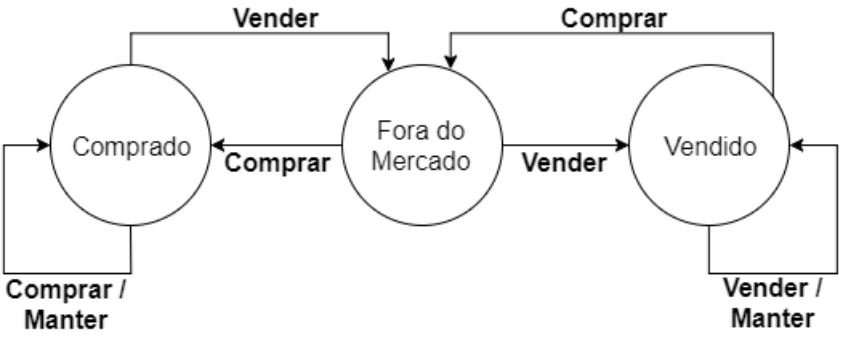

Figura 8 – Dinâmica das ações tomadas e posições do agente

4.3 Codificação da Recompensa

Segundo (SUTTON; BARTO, 1998), o objetivo de um agente em um problema de aprendizado por reforço é aprender uma política que maximize a sua recompensa obtida, decorrente da interação com o ambiente. Desse modo, a função de recompensa deve representar, de alguma forma, o objetivo final do agente.

Uma das funções de recompensa mais intuitivas para um problema financeiro é o retorno obtido no período. Como o intervalo de tempo considerado é de um dia, o retorno diário (Eq. (2.1)) é uma opção válida para a função de recompensa.

Para avaliar o desempenho de um agente, é preciso ter algum critério de comparação, um *benchmark*. Neste trabalho, o *benchmark* utilizado será o retorno da ação em questão na estratégia de *Buy and Hold* (B&H). Essa consiste em comprar o ativo e permanecer na mesma posição durante todo período considerado, vendendo-o no último dia.

Logo, o objetivo é obter um resultado melhor ou igual ao desempenho do *benchmark*. Para tal, modificou-se a fórmula do retorno diário para incorporar um maior peso às ações que gerem um resultado contrário ao desempenho do B&H, representado pelo retorno diário. Finalmente, a função de recompensa é mostrada a seguir:

$$R_t = \frac{1.5(P_t^a - P_{t-1}^a)}{P_{t-1}^a} - \frac{(P_t^b - P_{t-1}^b)}{P_{t-1}^b}, \tag{4.1}$$

em que P_t^a é o valor do *portfolio* e P_t^b é o preço da ação, ambos no instante de tempo t. O *portfolio* é definido como o número de ações em custódia do agente multiplicado pelo valor unitário da mesma, totalizando o capital investido no ativo.

A equação 4.1 consegue capturar algumas particularidades que podem auxiliar no treinamento do agente. Exemplificando, se o agente estiver em uma posição fora do mercado, não haverá mudança no valor do *portfolio* de um dia para o outro ($P_t^a - P_{t-1}^a = 0$). Entretanto, mesmo sem variação no *portfolio*, ele receberá uma recompensa positiva se o preço do ativo cair ($P_t^b - P_{t-1}^b < 0$) ou uma recompensa negativa se o mesmo subir ($P_t^b - P_{t-1}^b > 0$). Esse fato, por exemplo, não é capturado se só o retorno diário for utilizado como função de recompensa.

4.4 Codificação dos Estados

O agente toma decisões de acordo com seu estado atual. Logo, é extremamente importante que esse estado forneça informações úteis, que serão utilizadas pelo agente para tomar suas ações da melhor forma possível.

Em (MOODY; SAFFELL, 1999), utilizaram-se dados financeiros e macroeconômicos para retratação dos estados. Já em (CORAZZA; BERTOLUZZO, 2014), o retorno diário dos últimos N dias foi usado, onde N é uma variável que podia variar entre 1 e 5.

Neste trabalho, optou-se pela representação dos estados utilizando indicadores técnicos, descritos na Seção 2.6. Assume-se, então, uma abordagem segundo a análise técnica, baseada na hipótese da existência de tendências no mercado e na sua tendência de voltar à média.

Como afirmado, um dos fatores que dificulta aferir a real eficiência da análise técnica é a maneira subjetiva de se analisar visualmente gráficos e tendências. Dessa forma, procurou-se uma maneira quantitativa e objetiva de representação desses indicadores, que será descrita a seguir.

4.4.1 Retorno Diário

O retorno diário indica, objetivamente, a variação de P_t em relação a P_{t-1}. Portanto, ele pode ser usado diretamente como um dos componentes do estado, visto que não há necessidade de quantificar esse indicador.

4.4.2 Bandas de Bollinger

As bandas de Bollinger são interpretadas graficamente, observando o preço atual em relação ao valor das bandas superior e inferior. Esse indicador foi implementado de uma maneira "normalizada", na qual a distância do preço atual e da média móvel é dividida pelo desvio padrão, indicado a seguir:

$$Bollinger = \frac{P_n - SMA_N}{\sigma_N}, \tag{4.2}$$

onde SMA é a média móvel e σ_N é o desvio padrão do preço da ação no período de N dias. Nesse caso, utilizou-se $N = 20$.

4.4.3 Oscilador Estocástico e Média Móvel Convergente e Divergente

Existem dois aspectos principais na interpretação do oscilador estocástico e da média convergente e divergente: identificar quando a linha do indicador toca com a linha de gatilho e em qual valor isso ocorre. Para determinar o momento em que ocorre o contato entre as duas linhas, utiliza-se a função sinal, definida como:

$$sgn(x) = \begin{cases} -1 & \text{se x} < 0 \\ 0 & \text{se x} = 0 \\ 1 & \text{se x} > 0 \end{cases}$$

Se for feita a diferença entre a linha do indicador e a linha do gatilho para todo o intervalo de tempo, a mudança do sinal do resultado de um dia para o outro indica que as linhas se cruzaram. Pode-se usar esse fato e a função sinal para criar uma maneira de representar o toque das curvas e sua amplitude.

Para o oscilador estocástico, tem-se:

$$OE = \%D \mid sgn(\%D_n - \%K_n) - sgn(\%D_{n-1} - \%K_{n-1}) \mid . \tag{4.3}$$

É possível ver que OE terá valores diferentes de zero somente quando houver o cruzamento entre as linhas $\%D$ e $\%K$.

Similarmente, o mesmo procedimento foi realizado para a média móvel convergente e divergente:

$$MACD = \%LM \mid sgn(\%LM_n - \%LS_n) - sgn(\%LM_{n-1} - \%LS_{n-1}) \mid . \tag{4.4}$$

O estado do agente é formado pelos quatro indicadores descritos nessa seção, a Fig. 9 ilustra alguns valores possíveis para os estados.

Bollinger	Retorno Diário	MACD	Oscilador Estocástico
1.09	0.008	0	0
0.924	-0.001	0	0
1.75	0.018	2.307	0.0847
1.57	-0.004	0	0

Figura 9 – Exemplo de estados do agente, valores retirados diretamente do simulador

Cada linha corresponde ao valor dos indicadores no dia em questão. Como o aprendizado é feito de forma *online*, pode-se pensar no vetor de estados no dia t como $\mathbf{X}(S_t) = [\mathbf{x_0}(S_t), \mathbf{x_1}(S_t), \mathbf{x_2}(S_t), \mathbf{x_3}(S_t)]$. Fica evidente na Fig. 9 que o espaço de estados é praticamente contínuo, tornando o armazenamento de todos os valores possíveis para $Q(s, a)$ em uma tabela inviável. Dessa forma, é preciso **generalizar** estados semelhantes.

4.5 Aproximação da Função Q

Generalizações a partir de exemplos já vistos tem sido estudado extensivamente em outras áreas, como no aprendizado supervisionado (HAYKIN et al., 2009). Essa forma de generalização é conhecida como aproximação de função, em que uma função busca criar uma aproximação adequada de outra função objetivo (a função Q, por exemplo), a partir de exemplos fornecidos dessa função objetivo.

Uma forma bastante utilizada para realizar a aproximação da função Q é a utilização de uma arquitetura linear, onde essa função é representada por uma combinação linear dos parâmetros do estado S_t. Esse fato é representado na Fig. 10, onde a função recebe o vetor de estados $\mathbf{X}(S_t)$ e retorna valores de $\widehat{Q}(S_t, a_i)\ \forall a_i \in \mathcal{A}$

Nesse trabalho, a função Q foi aproximada por meio do algoritmo conhecido como LMS (do inglês, *Least Mean Squares*). Mais informações sobre o desenvolvimento e origem desse algoritmo podem ser encontradas em (HAYKIN; WIDROW, 2003).

A ideia geral do LMS é atualizar iterativamente os pesos $\mathbf{W}$ da arquitetura de forma a convergir para o conjunto de pesos ótimos, ou seja, que minimizem o erro quadrático médio. Os pesos são inicializados com pequenos valores e, a cada iteração, são

atualizados utilizando o gradiente descendente estocástico. As equações para a atualização dos pesos na enésima iteração são:

$$\mathbf{e}[n] = d[n] - \mathbf{W}x[n]; \tag{4.5}$$

$$\begin{aligned} \mathbf{W_{n+1}} &= \mathbf{W_n} + \Delta\mathbf{W} \\ &= \mathbf{W_n} + \mu\mathbf{e}[n]x[n], \end{aligned} \tag{4.6}$$

em que $x[n]$ são os dados de entrada, μ é a constante de aprendizado, $d[n]$ é a saída desejada e $\mathbf{W}x[n]$ é a saída estimada $\widehat{y}[n]$.

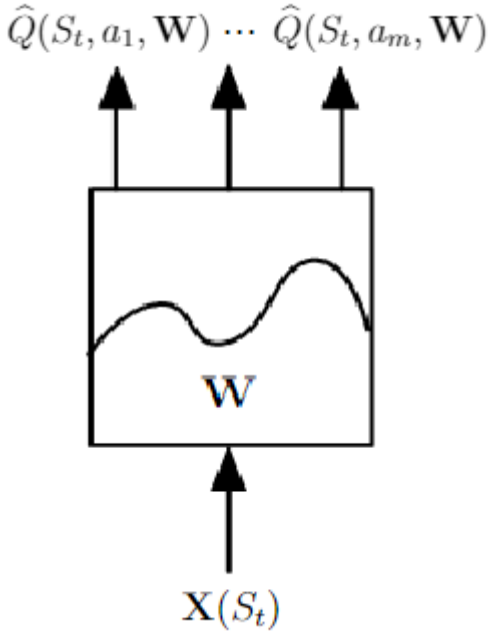

Figura 10 – Aproximação da Função Q

Quer-se aproximar $\widehat{Q}(S, A, \mathbf{W}) \approx Q_\pi(S, A)$, minimizando o erro quadrático médio da função estimada e do valor verdadeiro. Ou seja, minimizar a equação 4.7:

$$J(\mathbf{W}) = \mathbf{E}[(Q_\pi(S, A) - \widehat{Q}(S, A, \mathbf{W}))^2]. \tag{4.7}$$

Levando em consideração a aproximação linear utilizando o LMS, tem-se:

$$\widehat{Q}(S, A, \mathbf{W}) = \mathbf{X}(S_t)\mathbf{W}. \tag{4.8}$$

A atualização do pesos é dada pelo gradiente estocástico:

$$\begin{aligned} \Delta \mathbf{W} &= \mu(Q_\pi(S,A) - \widehat{Q}(S,A,\mathbf{W}))\nabla_{\mathbf{W}}\widehat{Q}(S,A,\mathbf{W}) \\ \nabla_{\mathbf{W}}\widehat{Q}(S,A,\mathbf{W}) &= \mathbf{X}(S_t) \end{aligned} \tag{4.9}$$

Na prática, o valor de $Q_\pi(S,A)$ não é conhecido (SILVER, 2015). Portanto, ele é substituído pela sua estimativa $R_{t+1} + \gamma\widehat{Q}(S_{t+1}, A_{t+1}, \mathbf{W})$:

$$\Delta \mathbf{W} = \mu(R_{t+1} + \gamma\widehat{Q}(S_{t+1}, A_{t+1}, \mathbf{W}) - \widehat{Q}(S,A,\mathbf{W}))\mathbf{X}(S_t) \tag{4.10}$$

Pode-se fazer um paralelo entre a Eq.(4.10) e Eq.(4.5) e Eq.(4.6). A diferença entre $R_{t+1} + \gamma\widehat{Q}(S_{t+1}, A_{t+1}, \mathbf{W}$ e $\widehat{Q}(S,A,\mathbf{W})$ é o sinal de erro e $\mathbf{X}(S_t)$ é a derivada da saída estimada.

4.6 Algoritmo

O algoritmo utilizado para o treinamento do agente é descrito no quadro a seguir. Entra-se com o número de épocas[1], o período de treinamento, o período de teste e qual o nome da ação[2].

A variável "DF" contém todas as informações da ação definida, como o número de ações em posse do agente, o valor do *portfolio*, o preço de fechamento, abertura, máximas e mínimas diárias. A variável "Estados" representa o conjunto de indicadores técnicos usados para todo o período de treinamento e o LMS é implementado conforme a descrição da seção anterior.

[1] Número de vezes que o algoritmo percorre todos os dados de treinamento
[2] Como a ação é referenciada na bolsa de valores.

Algorithm 1 *Algoritmo de Treinamento*

Require: $num\_epocas, periodo\_teste, periodo\_treinamento, nome\_ativo$

Inicializa DF
Inicializa $Estados$ e LMS com $\mathbf{W}$ arbitrário
for $(k = 0$ *to* $num\_epocas)$ **do**
  $t = 1$ {Volta ao inicio do período de treinamento}
  $S_t = Estados[t]$
  while $(periodo\_treinamento > t)$ **do**
    $\widehat{Q}(S_t, A_t) = S_t\mathbf{W}^T$
    if ϵ **then**
      Escolhe $a \in \mathcal{A}$ aleatoriamente
    else
      $a = argmax(\widehat{Q}(S_t, A_t))$
    end if
    Atualiza os valores de DF de acordo com a ação tomada
    Recebe R_{t+1} e S_{t+1}
    $\widehat{Q}(S_{t+1}, A_{t+1}) = S_{t+1}\mathbf{W}^T$
    $Q(S_t, a) \leftarrow Q(S_t, a) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, A_{t+1}) - Q(S_t, a)]$
    Atualiza $\mathbf{W}$ com o par S_t e $Q(S_t, A_t)$, por meio da Eq.(4.10)
    Incrementa t, decrementa ϵ
  end while
end for
Retorna $\mathbf{W}$

5 Resultados das Simulações

É indispensável considerar a temporalidade do problema ao se trabalhar com séries financeiras. Os dados financeiros de uma ação na década de 80, por exemplo, podem não conter praticamente nenhuma informação relevante para determinação do seu preço atual.

Não há nenhuma garantia que uma estratégia que funcionou bem no passado funcionaria atualmente. Por esse motivo, mantiveram-se os testes no período de tempo mais atual possível, de forma que os dados utilizados ainda tenham certa relevância nos dias atuais. Trabalhos como (ZANIN, 2008), (BROCKWELL; DAVIS, 2013) e (HAMILTON, 1994) trazem mais detalhes sobre técnicas de análise de séries temporais.

5.1 Testes com 6 Meses de Treinamento

O primeiro conjunto de resultados apresentado nesse trabalho foi obtido da seguinte forma: o agente foi submetido a treinamento por um período de 6 meses e, nos 3 meses subsequentes, foi aferido seu desempenho.

Utilizou-se essa metodologia para dados do começo de 2013 até Outubro de 2017. A ideia principal era testar o agente em um período imediatamente após o seu treinamento. Desse modo, acredita-se que o treinamento terá uma correlação maior com os dados de teste. Ao todo, 9 períodos distintos de teste e treinamento foram considerados.

Para essa rodada de testes, foram consideradas 3 ações da Bolsa de Valores de Nova Iorque ou *New York Stock Exchange* (NYSE). São elas:

- AAPL, referente à empresa *Apple Inc.*;
- IBM, referente à empresa *International Business Machines*;
- SPY, referente ao índice S&P 500[1].

Cada período de tempo foi simulado 10 vezes, totalizando 90 simulações para cada ação. O treinamento foi realizado utilizando 1000 épocas e a métrica escolhida para analisar o desempenho foi o de operações vitoriosas (O.V.). Operação vitoriosa, nesse caso, é quando o desempenho do agente é superior ao desempenho do B&H no mesmo período. A taxa de aprendizado utilizada foi de $0,1$ e a de desconto foi de $0,95$. A política adotada foi do tipo ϵ-ganaciosa, em que ϵ era iniciado em 1 e decrementado por $\frac{1}{num\_epocas}$ a cada episódio completado.

[1] Índice correspondente às ações das 500 maiores empresas norte-americanas cotadas na NYSE

Os resultados das simulações são apresentados na Tabela 3.

Período de Treinamento	**Período de Teste**	**O.V. AAPL**	**O.V. SPY**	**O.V. IBM**
10/01/2013 - 10/07/2013	10/07/2013 - 10/10/2013	10%	0%	90%
10/07/2013 - 10/01/2014	10/01/2014 - 10/04/2014	0%	20%	0%
10/01/2014 - 10/07/2014	10/07/2014 - 10/10/2014	10%	0%	80%
10/07/2014 - 10/01/2015	10/01/2015 - 10/04/2015	80%	100%	20%
10/01/2015 - 10/07/2015	10/07/2015 - 10/10/2015	20%	0%	100%
10/07/2015 - 10/01/2016	10/01/2016 - 10/04/2016	0%	90%	0%
10/01/2016 - 10/07/2016	10/07/2016 - 10/10/2016	0%	80%	80%
10/07/2016 - 10/01/2017	10/01/2017 - 10/04/2017	0%	90%	100%
10/01/2017 - 10/07/2017	10/07/2017 - 10/10/2017	90%	10%	0%
-	**Total**	23.33%	44.44%	52.2%

Tabela 3 – Porcentagens de Operações Vitoriosas

Em geral, os resultados não foram satisfatórios, principalmente para as ações da *Apple*. Um das principais hipóteses do desempenho fraco pode-se dever ao fato do treinamento ser realizado em um período de tempo muito curto. Desse modo, o agente fica limitado, muitas vezes, ao tipo de cenário econômico que ele encontra.

Exemplificando, no período de teste com início 10/01/2014, a AAPL teve um retorno de -2.3%. Entretanto, no período de treino, essa mesma ação obteve um retorno de 25.8%. Ou seja, o agente foi treinado em um cenário extremamente "otimista" e depois testado em um cenário onde a ação terminou o período com desempenho negativo.

Esse fato ocorre frequentemente, tanto para AAPL, quanto para as outras duas ações. É improvável que o agente tenha um bom desempenho se a natureza dos dados de seu treinamento diferem bastante dos dados do treino.

Dessa forma, procurou-se realizar o treinamento em um período de tempo maior, que forneça ao agente mais informações ao longo do treino. A seção a seguir descreve a nova rodada de simulações.

5.2 Testes com 4 Anos de Treinamento

Tendo em vista os problemas encontrados nos testes anteriores, os agentes nesta seção foram treinados durante um período maior de tempo.

O agente foi treinado do começo de 2013 até o final de 2016, totalizando 4 anos de treinamento e utilizando os mesmos valores de constante de aprendizagem e ϵ da seção anterior. O desempenho desse foi avaliado considerando diversos intervalos de tempo em 2017. Como o número de casos avaliados foi menor que o anterior, em vez de usar o número de operações vencedoras, utilizou-se a **diferença** do retorno obtido pelo algoritmo e do desempenho da ação, seguindo o B&H.

Vale ressaltar a importância da temporalidade nesse problema. O período de treinamento poderia ser mais longo, assim como o período de teste. Todavia, os resultados mais expressivos são os obtidos no tempo mais recente.

Além das 3 ações consideradas nas simulações anteriores, outras 5 foram acrescentadas para essa rodada de testes, escolhidas com base em dois aspectos principais: o volume negociado diariamente[2] e o tamanho da empresa. São elas:

- MSFT, referente à empresa *Microsoft Corporation*;
- FORD, referente à empresa *Ford Motor Company*;
- COCA-COLA, referente à empresa *The Coca-Cola Company*;
- CITI, referente à empresa *Citigroup Inc.*;
- FB, referente à empresa *Facebook*.

Os primeiros testes foram realizados ao longo dos meses de Janeiro a Setembro de 2017, da seguinte forma: o agente, já treinado, operava cada mês separadamente. Ou seja, o mês no qual o teste era realizado não sofria influência de operações consolidadas em outros meses. Os resultados dessas simulações são apresentados na Tabela 4.

Período de Teste	**AAPL**	**MSFT**	**SPY**	**IBM**	**FORD**	**COCA-COLA**	**CITI**	**FB**
Janeiro de 2017	-1.00%	**0.10%**	-0.22%	-0.58%	**3.94%**	**1.55%**	**0.53%**	0.00%
Fevereiro de 2017	-0.90%	-0.80%	-1.42%	**3.02%**	**1.12%**	**1.68%**	-2.20%	**2.26%**
Março de 2017	-0.08%	-0.10%	**0.21%**	**5.21%**	**0.10%**	-0.33%	**0.46%**	-1.60%
Abril de 2017	0.00%	0.00%	**0.67%**	**22.60%**	-1.30%	-0.18%	-0.59%	-2.00%
Maio de 2017	-0.80%	-0.24%	-0.19%	-2.20%	**0.90%**	-4.33%	-0.40%	-3.77%
Junho de 2017	0.00%	0.00%	-0.76%	-1.60%	-0.50%	**1.50%**	0.00%	-4.30%
Julho de 2017	**0.45%**	-0.95%	0.00%	**11.20%**	**0.95%**	**1.62%**	**1.00%**	-0.30%
Agosto de 2017	-0.40%	-0.13%	**0.10%**	-1.97%	0.00%	**1.24%**	**4.80%**	-1.80%
Setembro de 2017	**0.50%**	**0.55%**	-0.24%	-2.50%	**1.05%**	**2.41%**	**0.10%**	-0.20%
Total	-2.23%	-1.57%	-1.85%	**33.18%**	**6.26%**	**5.16%**	**2.64%**	-11.71%

Tabela 4 – Resultado das operações mês a mês

Similarmente, o agente operou em intervalos de 3 em 3 meses, como pode ser visto na Tabela 5 e de 5 em 5 meses, apresentado na Tabela 6.

É possível observar que, de modo geral, os resultados foram mais satisfatórios que os simulados no cenário de 6 meses de treinamento e 3 meses de teste.

Metade das ações tiverem resultados superiores ao B&H e, com exceção do FB, as outras ações ficaram com um desempenho relativamente próximo, ainda que inferior.

[2] Volume negociado é a quantidade total de dinheiro que a ação em questão movimenta, tanto em operações de compra como de venda. Como o volume das empresas consideradas é grande, pode-se assumir que a compra e venda de ações não impactará o preço da mesma

Período de Teste	AAPL	MSFT	SPY	IBM	FORD	COCA-COLA	CITI	FB
Jan - Abr de 2017	-2.16%	-0.85%	-2.06%	**9.60%**	**2.20%**	**0.84%**	-2.30%	**0.06%**
Abr - Jul de 2017	**1.60%**	**0.86%**	-1.02%	-0.53%	-1.30%	-2.52%	-0.50%	-9.80%
Jul - Out de 2017	-0.20%	-0.90%	-1.18%	-3.40%	**0.34%**	**2.41%**	**4.50%**	-3.03%
Total	-0.76%	-0.89%	-4.26%	**5.67%**	**1.24%**	**0.73%**	**1.70%**	-12.77%

Tabela 5 – Resultado das operações em intervalos de 3 em 3 meses

Período de Teste	AAPL	MSFT	SPY	IBM	FORD	COCA-COLA	CITI	FB
Jan - Maio de 2017	-1.33%	-0.89%	-1.46%	**32.00%**	**0.90%**	**0.30%**	-0.13%	-2.18%
Maio - Oct de 2017	-0.29%	**0.04%**	-2.01%	**2.25%**	**2.76%**	**4.38%**	**0.40%**	-13.90%
Total	-1.62%	-0.85%	-3.47%	**34.25%**	**3.66%**	**4.68%**	**0.27%**	-16.08%

Tabela 6 – Resultado das operações em intervalos de 5 em 5 meses

As últimas simulações realizadas consideram os meses de Janeiro a Outubro de 2017 como um único período de teste, o agente operou continuamente durante os meses em questão.

As Figuras 11 e 12 apresentam o desempenho do agente, da ação testada e o histórico das ações tomadas durante o período, segundo a codificação descrita no Capítulo 4. O eixo y das figuras corresponde ao retorno acumulado[3] e o eixo x corresponde ao tempo.

Figura 11 – Operações na *Apple*

[3] Os valores foram normalizados. Dessa forma, pode-se comparar o preço da ação com o retorno obtido pelo agente ao longo do período

Figura 12 – Operações na *Microsoft*

Pode-se observar certa semelhança no comportamento do agente ao operar as ações da *Apple* (Fig. 11) e da *Microsoft* (Fig. 12). Há uma forte tendência na manutenção de uma posição compradora, com algumas vendas pontuais.

A estratégia adotada pelo agente pode ser compreendida ao observar o treinamento para as duas ações em questão, apresentado no Anexo A. Nota-se que as operações de compra foram muito mais frequentes que as de venda. Tendo em vista que ambos ativos apresentaram grande crescimento no período, a estratégia adotada é coerente.

Similarmente, o agente optou pela continuidade de posições compradas na operação da *SPY* (Fig. 13), com algumas vendas momentâneas. Entretanto, seu desempenho foi inferior ao dos casos anteriores. As vendas realizadas em Março e em Setembro prejudicaram o algoritmo e o afastaram do desempenho da SPY.

É possível observar na Fig. 14, referente à *Ford*, a predominância das ordens de compra. Entretanto, as saídas dessas posições foram feitas em momentos mais adequados do que os casos anteriores. Esse fato pode ser observado pelo resultado ao final do teste, cujo é superior ao B&H.

Para as operações nos ativos da *Coca-Cola* e do *Citigroup*, nota-se uma maior diversidade nas ações tomadas pelo agente.

Na Fig. 15 alguns pontos merecem destaque: o agente fica fora do mercado nos

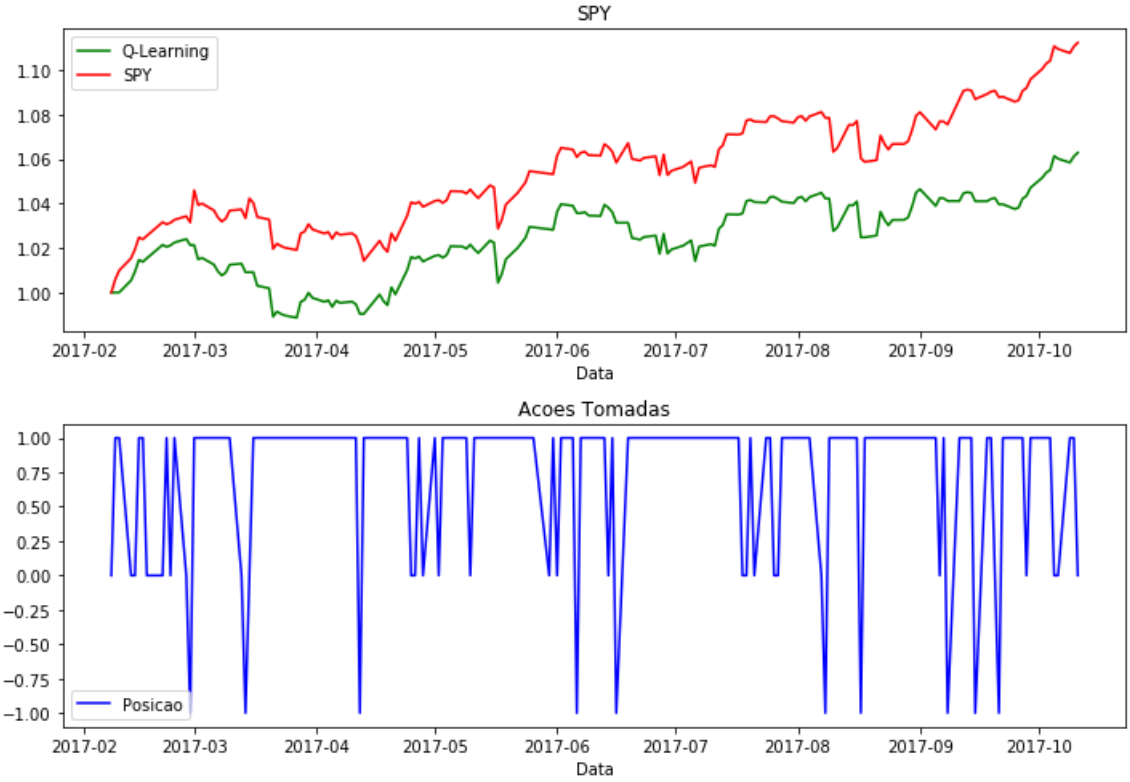

Figura 13 – Operações no *SP 500*

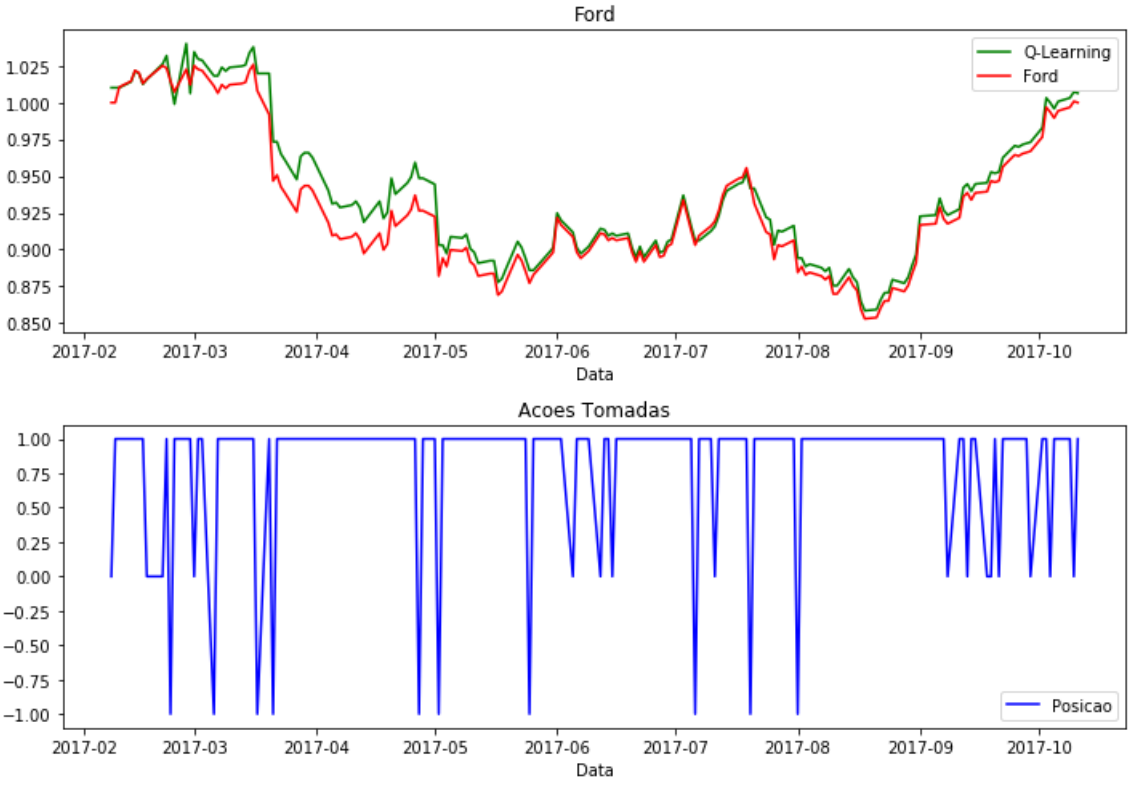

Figura 14 – Operações na Ford

primeiros dias, evitando uma queda brusca no preço, e, no meio de Maio, realiza um *short* em uma subida rápida de preços, prejudicando seu desempenho. As operações realizadas no restante do período conseguiram recuperar as perdas ocorridas por essa operação, resultando em um desempenho superior ao B&H.

Um cenário semelhante foi encontrado na Fig. 16. Contudo, o resultado foi o contrário, o agente perdeu a oportunidade de aproveitar um momento de subida no começo do período, ficando abaixo do desempenho da ação. Esse fato é compensado no meio de Março, onde um *short* leva o agente a superar a linha de *benchmark*.

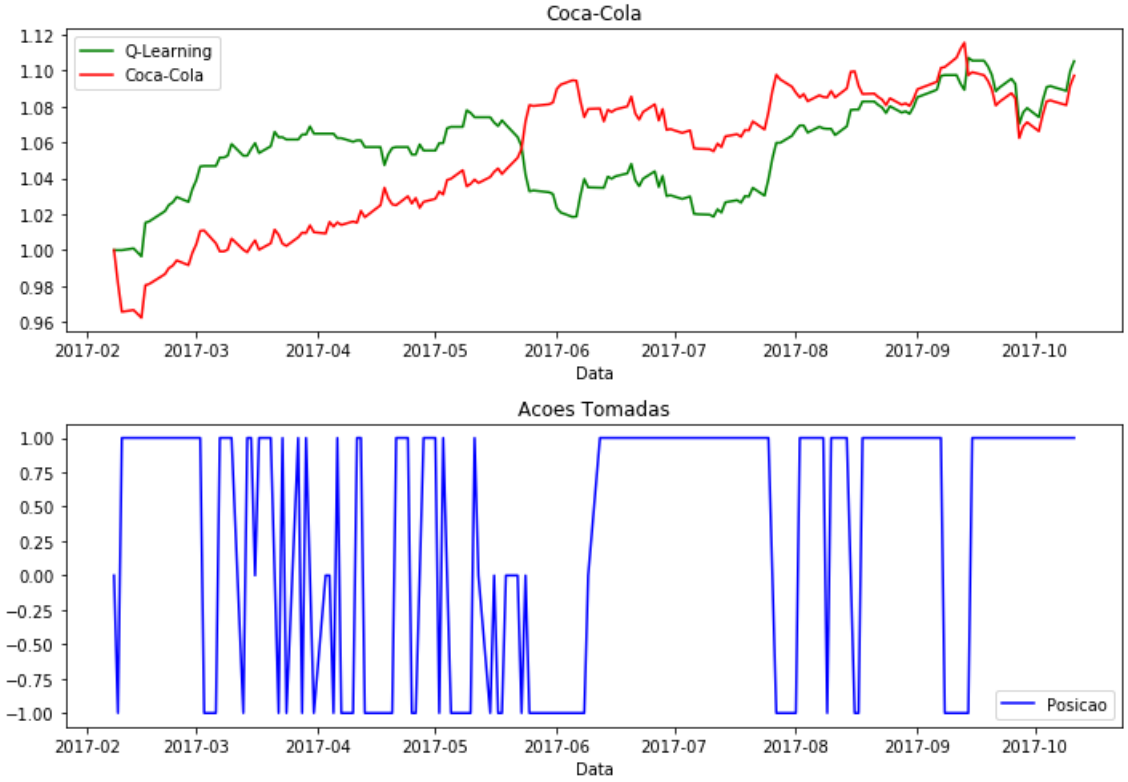

Figura 15 – Operações na *Coca-Cola*

Como observado nos testes anteriores, as ações da IBM apresentaram os melhores resultados. Esse acontecimento se repete nessa rodada de testes, como pode ser visto na Fig. 17. O agente realizou operações de *short* na maior parte do período, acarretando em um desempenho bem superior ao da IBM, que apresentou quedas frequentes.

Por fim, a Fig. 18 apresenta os resultados do agente operando FB. Embora essa apresente o pior desempenho dentre as ações testadas, é proveitoso observar alguns eventos que ocorreram durante o teste.

No meio de Maio, a ação sofre uma queda considerável, que é interpretada como uma oportunidade para *short*. Entretanto, o ativo continua a subir. Esse fato ocorre novamente no meio de Junho.

Figura 16 – Operações no *Citigroup*

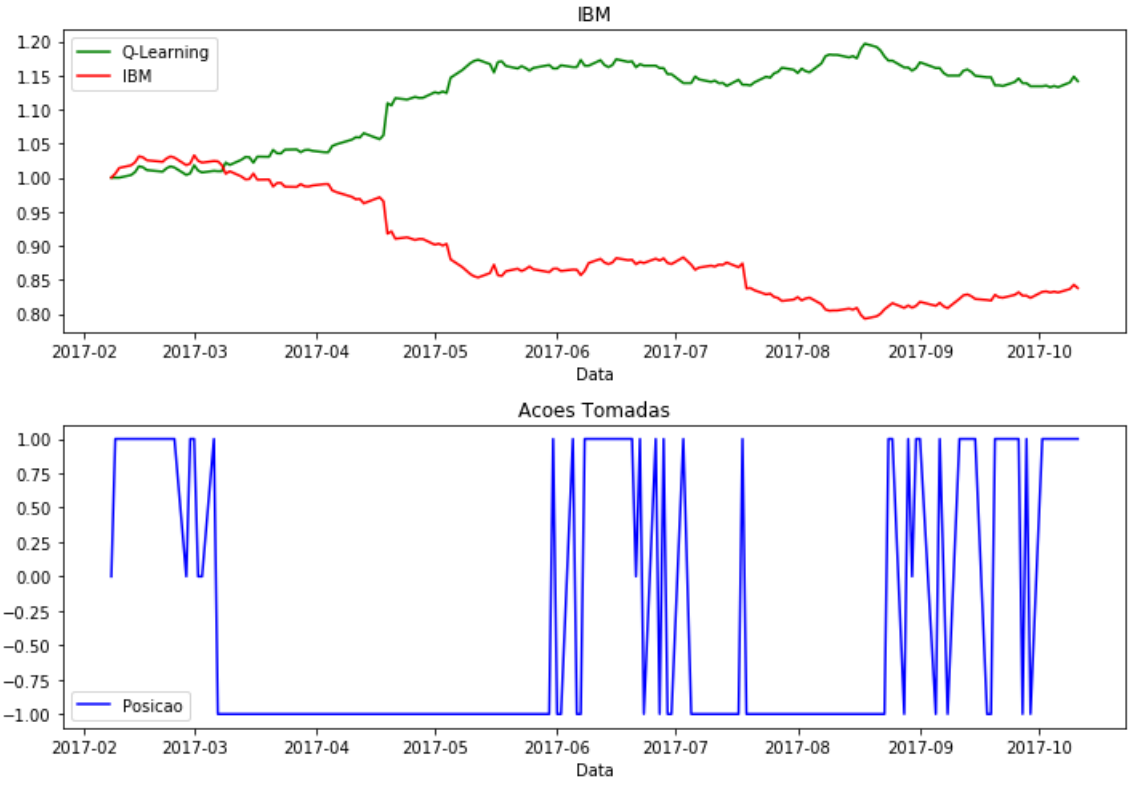

Figura 17 – Operações na IBM

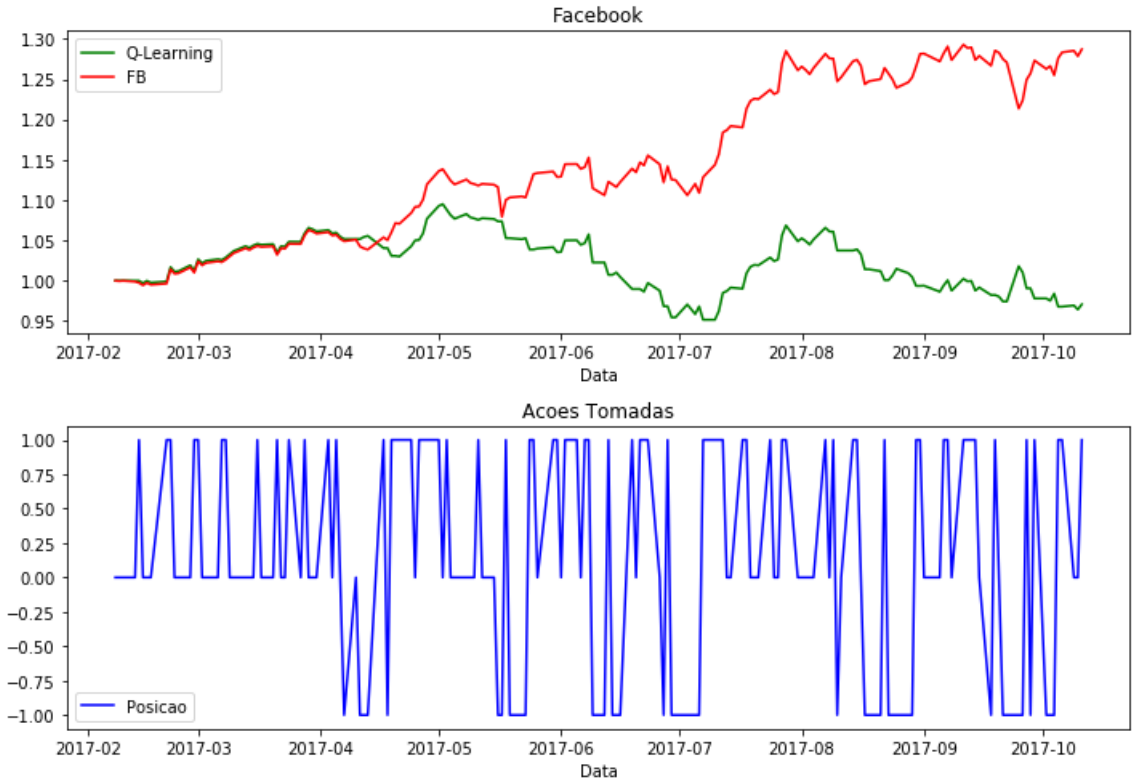

Figura 18 – Operações no *Facebook*

5.3 Comentários Gerais

Em uma primeira análise, tende-se a concluir que o algoritmo utilizado apresenta resultados melhores que o B&H em algumas ações e resultados piores em outras, porém, sem um padrão definido. Entretanto, essa conclusão pode ser, de certa forma, simplista.

O modelo para *trading* desenvolvido neste trabalho utiliza indicadores técnicos para a tomada de decisões. Como já elucidado, a análise técnica apresenta algumas limitações, visto que a mesma está embasada em uma série de hipóteses e só faz uso de dados de preço e volume do ativo sendo analisado.

Para a melhor interpretação dos resultados, é fundamental atentar-se às empresas por trás dos ativos.

As três empresas cujas ações não foram superadas pelo *Q-Learning* (AAPL,FB e MSFT) são do ramo de tecnologia, segmento que tem crescido bastante nos últimos anos. Isto posto, diversos analistas e jornais classificam o cenário econômico atual como uma possível "bolha tecnológica"(*Tech Bubble*), por conta do seu alto crescimento (SHARMA, 2017).

Os principais ativos do ramo de tecnologia dos Estados Unidos tiveram um desempenho, em média, 40% maior que o da S&P 500 (RICH, 2017). Desse modo, torna-se difícil para o agente obter resultados expressivos, visto que o mesmo utiliza parâmetros

que partem de pressupostos como Retorno à Média e outras características que não são observadas nessas ações.

O outro cenário que apresentou desempenho inferior ao B&H foram as operações no SPY, referente ao índice S&P 500. O preço dessa ação é ponderado pelo valor de mercado de cada uma das 500 empresas por ele representadas. Efetivamente, esse índice contém informações de todas as empresas participantes. Novamente, o agente encontra dificuldade em utilizar indicadores técnicos em um ativo que apresenta um movimento de preço bem mais complexo, formado por 500 variações ponderadas.

No entanto, obteve-se bons resultados nas ações de outros setores da economia, como financeiro (*Citi*), bebidas (*Coca-Cola*) e automobilístico (*Ford*). Em geral, as ações de empresas desses setores apresentam um comportamento "esperado"na visão da análise técnica. É importante ressaltar que isso não diz respeito à previsibilidade do movimento das ações, mas sim da variação de seus preços, que não apresentaram crescimentos bem acima da média, como as do ramo de tecnologia.

Por fim, o melhor desempenho, quando comparado ao *benchmark*, foi obtido operando IBM, uma empresa do ramo de tecnologia. Diferente das outras empresas desse ramo testadas, a IBM não vem apresentando resultados expressivos de receita e lucro (SCHOFIELD, 2017).Consequentemente, o preço de sua ação vem sofrendo diversas perdas nos últimos anos (MATERRA, 2015). Desse modo, o algoritmo pode obter bons resultados, realizando diversas operações de venda no período de baixa do ativo.

6 Conclusões

Neste trabalho, buscou-se apresentar um modelo capaz de desenvolver uma estratégia de negociação de ações baseado no paradigma de aprendizado por reforço.

Em RL, o agente não é informado quais decisões devem ser tomadas, este aprende por meio da interação direta com o ambiente, descobrindo a maneira ótima de agir por meio de tentativa e erro. Para tal, um ambiente de simulações foi desenvolvido para que o agente possa atuar, observar recompensa e aprender, utilizando o algoritmo *Q-Learning*.

A representação dos estados é fundamental em um problema de aprendizado por reforço, e é um ponto que merece ser destacado. Os estados foram codificados utilizando quatro indicadores técnicos. Devido ao grande número de estados possíveis, optou-se pela utilização de um função linear para generalizar a função Q. Apesar das críticas existentes a respeito da utilização da análise técnica, os resultados foram satisfatórios, a estratégia B&H foi superada em metade dos casos. Não obstante, as limitações existentes nessa metodologia foram observadas nas simulações realizadas e discutidas no texto.

O efeito do cenário de treinamento é outro aspecto interessante notado neste trabalho. Quando o agente era treinado em um período com fortes tendências de alta ou baixa, seu desempenho no teste poderia ser medíocre. Um intervalo de tempo maior para treinamento resultou em desempenhos superiores.

Por fim, conforme apontado na Introdução, ainda há bastante espaço para a aplicação de técnicas de inteligência artificial no contexto de investimentos e, consequentemente, possíveis extensões para esse trabalho. A primeira seria acrescentar mais indicadores na representação dos estados, como indicadores de volume e outras formas de representação gráfica, e até indicadores macroeconômicos.

Outra possível alternativa seria a mudança da função de aproximação, em vez de utilizar uma arquitetura linear pode-se usar uma árvore de decisões ou rede neural artificial para aproximar a função Q. Ou ainda, outra função de recompensa poderia ser utilizada que incorporasse os custos de transação existentes na bolsa de valores.

Finalmente, outra possível extensão seria a utilização de outro algoritmo de RL, como o *recurrent reinforcemente learning*, testado por (DU; ZHAI; LV, 2016) e a utilização de uma variedade de ativos diferentes para a validação do modelo, inclusive de empresas listadas na Bovespa.

Referências

ALDRIGHI, D. M.; MILANEZ, D. Y. Finança comportamental e a hipótese dos mercados eficientes. *Revista de Economia Contemporânea*, v. 9, n. 1, p. 41–72, 2005. Citado na página 29.

APPEL, G. *The moving average convergence-divergence trading method: advanced version*. [S.l.]: Scientific Investment Systems, 1985. Citado na página 36.

BACHELIER, L. *Théorie de la spéculation*. [S.l.]: Gauthier-Villars, 1900. Citado na página 28.

BLACK, F. Noise. *The journal of finance*, Wiley Online Library, v. 41, n. 3, p. 528–543, 1986. Citado na página 29.

BRITO, B.; ARAUJO, D. Aprendizado por reforco aplicado ao mercado financeiro. In: . [S.l.: s.n.], 2013. Citado na página 22.

BROCKWELL, P. J.; DAVIS, R. A. *Time series: theory and methods*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 55.

CHAND, S.; CHANDRA, R. Cooperative coevolution of feed forward neural networks for financial time series problem. In: IEEE. *Neural Networks (IJCNN), 2014 International Joint Conference on*. [S.l.], 2014. p. 202–209. Citado na página 22.

CHAO-WEN, L.; JR, M. R. R. Ewma control charts for monitoring the mean of autocorrelated processes. *Journal of Quality technology*, American Society for Quality, v. 31, n. 2, p. 166, 1999. Citado na página 33.

COHEN, K. J. et al. Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics*, Elsevier, v. 12, n. 2, p. 263–278, 1983. Citado na página 28.

CORAZZA, M.; BERTOLUZZO, F. Q-learning-based financial trading systems with applications. *University Ca' Foscari of Venice, Dept. of Economics Working Paper Series*, 2014. Citado 2 vezes nas páginas 22 e 49.

DAVIDSON, R. *Trading for a Living*. [S.l.]: Penton Overseas, Incorporated, 2000. Citado na página 30.

DU, X.; ZHAI, J.; LV, K. Algorithm trading using q-learning and recurrent reinforcement learning. *Journal Positions*, v. 1, p. 1, 2016. Citado na página 65.

ELDER, A. *Come into my trading room: A complete guide to trading*. [S.l.]: John Wiley & Sons, 2002. v. 163. Citado 3 vezes nas páginas 26, 27 e 30.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970. Citado na página 28.

FRENCH, K. R. Permanent and temporary components of stock prices. *Journal of political Economy*, The University of Chicago Press, v. 96, n. 2, p. 246–273, 1988. Citado na página 28.

GRAHAM, B. *The intelligent investor: A book of practical counsel*. [S.l.]: Prabhat Prakashan, 1965. Citado na página 29.

GRANVILLE, J. E. *New key to stock market profits*. [S.l.]: Prentice-Hall, 1963. Citado 2 vezes nas páginas 25 e 27.

GROSSMAN, S. J.; STIGLITZ, J. E. On the impossibility of informationally efficient markets. *The American economic review*, JSTOR, v. 70, n. 3, p. 393–408, 1980. Citado na página 29.

HALL, T.; KUMAR, N. *Why Machine Learning Models Often Fail to Learn*. 2016. Disponível em: <https://www.bloomberg.com/news/articles/2016-11-10/why-machine-learning-models-often-fail-to-learn-quicktake-q-a>. Citado na página 22.

HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton university press Princeton, 1994. v. 2. Citado na página 55.

HAYKIN, S.; WIDROW, B. *Least-mean-square adaptive filters*. [S.l.]: John Wiley & Sons, 2003. v. 31. Citado na página 51.

HAYKIN, S. S. et al. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, NJ, USA:, 2009. v. 3. Citado na página 51.

INVESTOPEDIA. *Investopedia*. 2017. Disponível em: <http://investopedia.com>. Citado 2 vezes nas páginas 15 e 36.

LO, A. W. The adaptive markets hypothesis. *The Journal of Portfolio Management*, Institutional Investor Journals, v. 30, n. 5, p. 15–29, 2004. Citado na página 29.

LO, A. W.; MACKINLAY, A. C. Stock market prices do not follow random walks: Evidence from a simple specification test. *The review of financial studies*, Oxford University Press, v. 1, n. 1, p. 41–66, 1988. Citado na página 28.

MATERRA, S. *IBM Used To Be Bigger Than Apple, What Happened?* 2015. Disponível em: <https://www.fool.com/investing/general/2015/04/11/ibm-used-to-be-bigger-than-apple-what-happened.aspx>. Citado na página 64.

MOODY, J.; WU, L. Optimization of trading systems and portfolios. In: IEEE. *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*. [S.l.], 1997. p. 300–307. Citado na página 22.

MOODY, J. E.; SAFFELL, M. Reinforcement learning for trading. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 1999. p. 917–923. Citado 3 vezes nas páginas 21, 22 e 49.

MURPHY, J. J. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. [S.l.]: Penguin, 1999. Citado na página 34.

NEVMYVAKA, Y.; FENG, Y.; KEARNS, M. Reinforcement learning for optimized trade execution. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.], 2006. p. 673–680. Citado na página 22.

PANELLA, M.; LIPARULO, L.; PROIETTI, A. A higher-order fuzzy neural network for modeling financial time series. In: IEEE. *Neural Networks (IJCNN), 2014 International Joint Conference on*. [S.l.], 2014. p. 3066–3073. Citado na página 22.

RICH, B. *Has The Tech Boom Run Its Course?* 2017. Disponível em: <https://www.forbes.com/sites/bryanrich/2017/09/27/faang-has-the-tech-boom-run-its-course/#59ba964658c1>. Citado na página 63.

SCHOFIELD, J. *IBM's five years of falling revenues have left it smaller than Apple, Microsoft and Google*. 2017. Disponível em: <http://www.zdnet.com/article/ibms-five-years-of-declining-revenues-have-left-it-smaller-than-apple-microsoft-and-google/>. Citado na página 64.

SHARMA, R. *When Will the Tech Bubble Burst?* 2017. Disponível em: <https://www.nytimes.com/2017/08/05/opinion/sunday/when-will-the-tech-bubble-burst.html>. Citado na página 63.

SILVER, D. *Lecture notes in Reinforcement Learning*. [S.l.]: University College London, 2015. Citado 3 vezes nas páginas 40, 42 e 53.

STOCKCHART. *StockChart*. 2017. Disponível em: <http://stockcharts.com>. Citado 2 vezes nas páginas 15 e 34.

SUTTON, R. S.; BARTO, A. G. *Reinforcement learning: An introduction*. [S.l.]: MIT press Cambridge, 1998. v. 1. Citado 4 vezes nas páginas 39, 42, 44 e 48.

VOHS, K. D.; MEAD, N. L.; GOODE, M. R. The psychological consequences of money. *science*, American Association for the Advancement of Science, v. 314, n. 5802, p. 1154–1156, 2006. Citado na página 21.

VOHS, K. D.; MEAD, N. L.; GOODE, M. R. Merely activating the concept of money changes personal and interpersonal behavior. *Current Directions in Psychological Science*, SAGE Publications Sage CA: Los Angeles, CA, v. 17, n. 3, p. 208–212, 2008. Citado na página 21.

WATKINS, C. J.; DAYAN, P. Q-learning. *Machine learning*, Springer, v. 8, n. 3-4, p. 279–292, 1992. Citado na página 46.

WATKINS, C. J. C. H. *Learning from delayed rewards*. Tese (Doutorado) — King's College, Cambridge, 1989. Citado na página 45.

ZANIN, M. Forbidden patterns in financial time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, AIP, v. 18, n. 1, p. 013119, 2008. Citado na página 55.

ZEMKE, S. On developing a financial prediction system: Pitfalls and possibilities. In: *Proceedings of the First International Workshop on Data Mining Lessons Learned (DMLL-2002)*. [S.l.: s.n.], 2002. p. 8–12. Citado na página 21.

Anexos

ANEXO A – Primeiro Anexo

Figuras referentes ao período de treinamento de AAPL e MSFT

Figura 19 – Período de Treinamento da AAPL

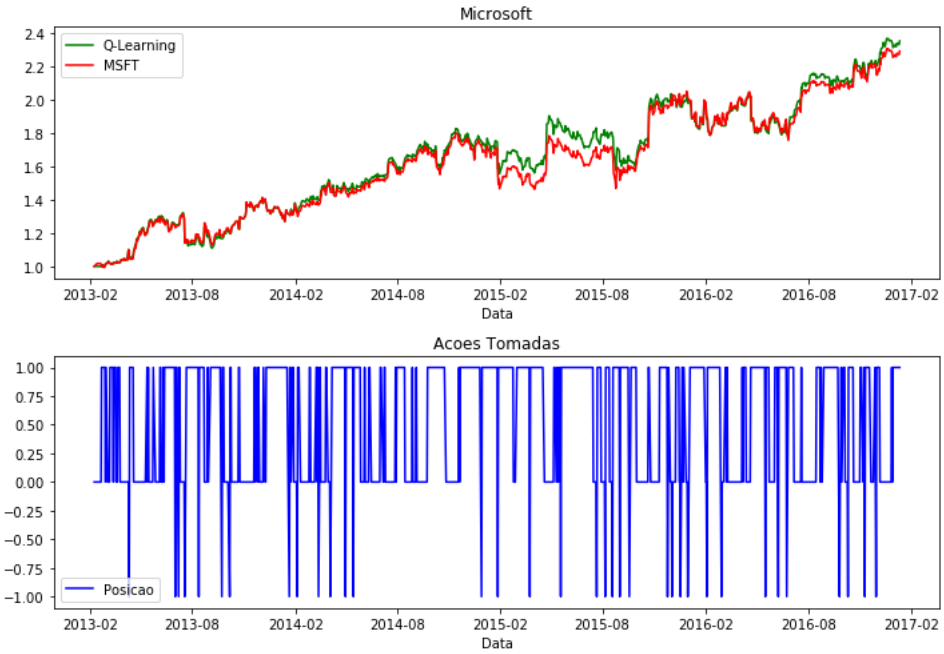

Figura 20 – Período de Treinamento da MSFT